From Secondary Structure to Three-Dimensional Structure: Improved Dihedral Angle Probability Distribution Function for Use with Energy Searches for Native Structures of Polypeptides and Proteins*

BETTY CHENG, AKBAR NAYEEM,† and HAROLD A. SCHERAGA‡

Baker Laboratory of Chemistry, Cornell University, Ithaca, New York, 14853-1301

Received 22 September 1995; accepted 29 November 1995

ABSTRACT

An improved scheme to help in the prediction of protein structure is presented. This procedure generates improved starting conformations of a protein suitable for energy minimization. Trivariate gaussian distribution functions for the ϕ, ψ, and χ^1 dihedral angles have been derived, using conformational data from high resolution protein structures selected from the Protein Data Bank (PDB). These trivariate probability functions generate initial values for the ϕ, ψ, and χ^1 dihedral angles which reflect the *experimental* values found in the PDB. These starting structures speed the search of the conformational space by focusing the search mainly in the regions of native proteins. The efficiency of the new trivariate probability distributions is demonstrated by comparing the results for the α-class polypeptide fragment, the mutant *Antennapedia* (C39 → S) homeodomain (2HOA), with those from two reference probability functions. The first reference probability function is a uniform or flat probability function and the second is a bivariate probability function for ϕ and ψ. The trivariate gaussian probability functions are shown to search the conformational space more efficiently than the other two probability functions. The trivariate gaussian probability functions are also tested on the binding domain of *Streptococcal*

*This article includes Supplementary Material available from the authors upon request or via the Internet at ftp.wiley.com/public/journals/jcc/suppmat/17/1453 or http://www.wiley.com/jcc

† Current address: Tripos, Inc., 1699 S. Hanley Road, Suite 303, St. Louis, MO, 63144

‡ Author to whom all correspondence should be addressed.

Journal of Computational Chemistry, Vol. 17, No. 12, 1453–1480 (1996)

CCC 0192-8651 / 96 / 121453-25

protein G (2GB1), an α/β class protein. Since presently available energy functions are not accurate enough to identify the most native-like energy-minimized structures, three selection criteria were used to identify a native-like structure with a 1.90-Å rmsd from the NMR structure as the best structure for the *Antennapedia* fragment. Each individual selection criterion (ECEPP/3 energy, ECEPP/3 energy-plus-free energy of hydration, or a knowledge-based mean field method) was unable to identify a native-like structure, but simultaneous application of more than one selection criterion resulted in a successful identification of a native-like structure for the *Antennapedia* fragment. In addition to these tests, structure predictions are made for the *Antennapedia* polypeptide, using a Pattern Recognition-based Importance-Sampling Minimization (PRISM) procedure to predict the backbone *conformational state* of the mutant *Antennapedia* homeodomain. The ten most probable backbone conformational state predictions were used with the trivariate and bivariate gaussian dihedral angle probability distributions to generate starting structures (i.e., dihedral angles) suitable for energy minimization. The final energy-minimized structures show that neither the trivariate nor the bivariate gaussian probability distributions are able to overcome the inaccuracies in the backbone conformational state predictions to produce a native-like structure. Until highly accurate predictions of the backbone conformational *states* become available, application of these dihedral angle probability distributions must be limited to problems, such as homology modeling, in which only a limited portion of the backbone (e.g., surface loops) must be explored.

Introduction

Various methods are being used to surmount the multiple-minima problem in computational methods to determine the native structures of proteins.[1] As an aid in these calculations, it is helpful to limit the region of conformational space that must be searched to locate the global minimum of the conformational energy. For this purpose, several authors,[2-6] have used the Brookhaven Protein Data Bank[7] (PDB) to obtain probability distribution functions for searching the conformational space of the backbone dihedral angles ϕ and ψ of the residues of an unknown protein.

To limit the conformational space to a greater degree, it is helpful to have trivariate probability distribution functions to restrict the ranges of the three dihedral angles ϕ, ψ, and χ^1. The first goal of this article is to describe the derivation of such trivariate probability distribution functions. Evans *et al.*[5] have also reported trivariate probability distribution functions but they adopted discrete rather than continuous values of ϕ, ψ, and χ^1. Many other studies have reported the conformational preferences of residues based on surveys of experimental structures[2-6, 8-19]; however, most of the applications have focused on side-chain conformations in the context of a fixed backbone conformation.

The second goal of this article is to evaluate the improvement obtained by using these trivariate probability distribution functions. Finally, a third goal is to use these distribution functions as the first step in energy minimization, in an initial attempt to predict the native structure of two proteins.

It is convenient to think of protein structure prediction in two stages. In the first stage, the secondary structure of the backbone must be identified. Instead of the normal definition of secondary structural states (helical, β-sheet, coil) we use the following definition. Four backbone *conformational states* (α, ϵ, α^*, and ϵ^*) are defined by dividing the ϕ, ψ map of each residue into four regions (see Fig. 1).[2-4] α-Helical structures are found in the region labeled α, β-sheet structures correspond to the ϵ region, etc. In the second stage, the three-dimensional structure (i.e., its dihedral angles), *given* the backbone structure (defined in terms of the *conformational state* of each residue), must be determined. Initially, we assume that the native *conformational state* of the backbone is available, and use the ϕ, ψ, and χ^1 probability distribution functions to generate starting conformations (dihedral angles) for subsequent energy minimization with the Empirical Conformational Energy Program for Peptides (ECEPP) algorithm.[20-23] Suitable initial values are chosen for all other side-chain dihedral angles beyond χ^1, as described in the Methods section.

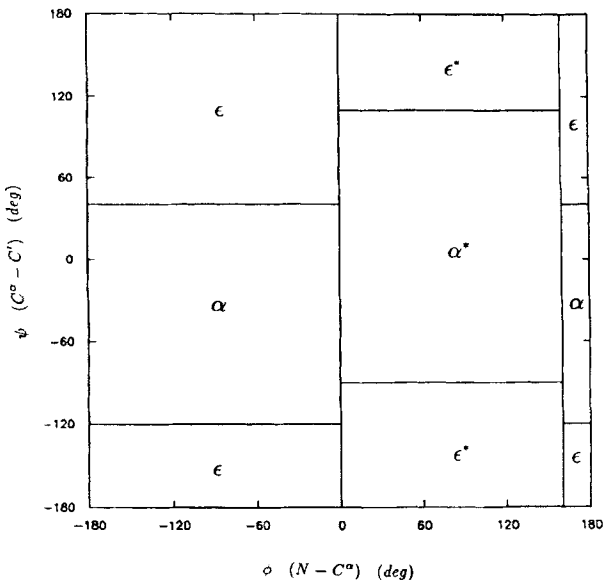

FIGURE 1. ϕ, ψ map showing the definitions of the four backbone conformational states adopted by Lambert and Scheraga.[2-4]

The performance of the ϕ, ψ, and χ^1 dihedral angle probability distribution functions is compared with the performance of two reference probability distribution functions. One of these is an unbiased or uniform probability distribution function which provides a baseline for reference, and the other is a bivariate probability distribution function for ϕ and ψ only,[2-4] thereby allowing us to evaluate the effect of including χ^1 in the trivariate function.

In the initial calculations, the backbone *conformational states* were adopted with 100% accuracy, and the various probability distribution functions were used to generate initial backbone and sidechain dihedral angles for energy-minimization calculations. In subsequent calculations, the backbone *conformational states* were obtained by prediction to evaluate the effect of inaccuracies in the backbone structure prediction on the performance of the probability distribution functions. In the latter computations, the bivariate and trivariate probability distribution functions were also used to generate starting dihedral angles from these predicted backbone conformational states. After selection of the starting dihedral angles, the energies of these starting structures were then minimized.

The efficiency with which the probability distribution functions explore the native regions of the conformational space was assessed by the root-mean-square deviations (rmsd) of the energy-minimized structures from the experimental structure. Examination of the lowest rmsd structure and the distribution of rmsd's for the energy-minimized structures revealed which of the probability distribution functions performed best.

We use three methods to identify the lowest rmsd structure: (a) identification of an ensemble of low-energy structures with ECEPP/3[23]; (b) inclusion of an empirical free energy of hydration term in the ECEPP/3 energies; and (c) use of the knowledge-based mean field method of Sippl and coworkers.[24-31]

Comparison of three different probability distribution functions (trivariate gaussian, bivariate gaussian, uniform) (with 100% accuracy assumed for the backbone *conformational states*) was carried out for the mutant *Antennapedia* (C39 → S) homeodomain (2HOA), which is the conserved domain of proteins implicated in homeotic transformations in which one part of an organism develops as the likeness of another part. Attempts were then made to predict the structures of two proteins, 2HOA, and the binding domain of *Streptococcal* protein G (2GB1) by using different energy criteria and a knowledge-based mean field method to identify native structures. To test the effect of a real, i.e., inaccurate, *conformational state* prediction for 2HOA, two types of computations were carried out (one with the assumption of 100% accuracy for the backbone *conformational states* and one with the backbone *conformational states* predicted by the **P**attern **R**ecognition-based **I**mportance-**S**ampling **M**inimization [PRISM] procedure[2-4]); for 2GB1, only computations with the assumption of 100% accuracy for the backbone *conformational states* were carried out. Their coordinates were not included in the PBD data set used to obtain the probability distribution functions. The coordinates of 2HOA and 2GB1 are available from the PDB, but were originally obtained from K. Wüthrich and G. M. Clore, respectively.

Methods

DATA SET

To obtain accurate probability distribution functions for the dihedral angles, high-resolution structures were selected from the 1989 release of the PDB.[7] The proteins and polypeptides chosen are listed in Table I of the supplementary materials.[32] The X-ray structures were of at least 2 Å resolution and had R-factors lower than 0.27. To reduce bias,

TABLE I.
Distributions of Conformational States in the (ϕ, ψ, χ^1) Space for the Cys[a] Residue.[b]

N[c]	Conf.[d]	μ_ϕ (deg.)	μ_ψ (deg.)	μ_{χ^1} (deg.)	$\Sigma_{\phi\phi}$ (deg.)^2	$\Sigma_{\phi\psi}$ (deg.)^2	$\Sigma_{\phi\chi^1}$ (deg.)^2	$\Sigma_{\psi\psi}$ (deg.)^2	$\Sigma_{\psi\chi^1}$ (deg.)^2	$\Sigma_{\chi^1\chi^1}$ (deg.)^2
78	αg^-	−76.96	−28.33	−64.27	473.02	194.63	−18.01	376.34	−18.78	109.67
16	αg^+	−77.13	−20.98	61.30	590.69	−355.62	−67.19	451.64	17.86	193.00
26	αt	−65.64	−43.45	−176.33	62.21	1.98	20.91	48.84	6.69	93.14
69	ϵg^-	−107.98	49.04	−65.48	635.30	−108.42	80.96	511.39	−67.00	172.60
19	ϵg^+	−132.80	159.67	63.71	962.39	−56.75	102.05	136.81	7.91	99.33
45	ϵt	−105.53	125.41	−176.94	1278.05	−123.35	−44.80	315.63	77.81	185.38
4	$\alpha^* g^-$	55.50	38.62	−51.20	99.77	−201.74	28.61	435.19	−110.46	126.05
0	$\alpha^* g^+$	***[e]	***	***	***	***	***	***	***	***
1	$\alpha^* t$	54.53	42.00	164.94	***	***	***	***	***	***
0	$\epsilon^* g^-$	***	***	***	***	***	***	***	***	***
0	$\epsilon^* g^+$	***	***	***	***	***	***	***	***	***
0	$\epsilon^* t$	***	***	***	***	***	***	***	***	***

[a] Both the cysteine (—SH) and the cystine (S—S) residues are counted in this table.
[b] The quantities μ_i and Σ_{ij} are defined in eqs. (7) and (8), respectively.
[c] N is the number of occurrences of the conformational state. For example, the Cys residue is found in the αg^- state 78 times in the data set.
[d] The 12 conformational states are defined in the "Description of Conformational States" section.
[e] *** has been inserted when the quantities in question cannot be calculated because of the absence of data.

sequences were screened for homology using the procedure of Needleman and Wunsch.[33] All proteins with sequence identities exceeding 50% with respect to previously accepted sequences were removed. The same criterion was applied to any homologies among chains for proteins with more than a single chain as well as any homologies that might exist within different segments of the same chain.

The experimental data form 12 clusters in the ϕ, ψ, and χ^1 space. Hence, 12 trivariate gaussian functions were used to fit the experimental data for each residue except for Gly, Ala, and Pro.

Glycine does not have a side chain and, therefore, it has no χ^1 dihedral angle. Since the g^-, g^+, and t positions are sterically identical for the alanine side chain, the value of χ^1 was always set to 60°. Proline has a fixed side chain and a fixed backbone dihedral angle, ϕ. Because the ECEPP force field uses rigid geometry, the pyrrolidine ring of proline was constrained to be in one of two positions.[20-23] ϕ was allowed to be either −68.8°, the "down" position, or −53.0°, the "up" position.[23] The bivariate probability function for proline generates values for ϕ and ψ, but the value of ϕ was reset to the closest allowed value, either "up" or "down." Because of the unique nature of the side chains for Gly, Ala, and Pro, only bivariate gaussian functions (in terms of ϕ and ψ) were used to fit the experimental data.

FITTING TRIVARIATE GAUSSIAN FUNCTIONS TO EXPERIMENTAL DIHEDRAL ANGLE DATA

A point, $\mathbf{x}$, in the ϕ, ψ, χ^1 space, may be represented as

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} \phi \\ \psi \\ \chi^1 \end{pmatrix} \tag{1}$$

where the values of ϕ, ψ, and χ^1 all lie between −180° and 180°. The trivariate gaussian probability function,[34] $f(\mathbf{x})$, is written as

$$f(\mathbf{x}) = \frac{1}{2\pi^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \tag{2}$$

where p is the dimensionality of the space; in the present work, p is equal to three (or two for the bivariate gaussian functions); $\boldsymbol{\mu}$ is the centroid of the distribution, $\boldsymbol{\Sigma}$ is the covariance matrix, and $|\boldsymbol{\Sigma}|$ is the determinant of the covariance matrix. The covariance matrix may be written in the following manner

$$\boldsymbol{\Sigma} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\ \Sigma_{21} & \Sigma_{22} & \Sigma_{23} \\ \Sigma_{31} & \Sigma_{32} & \Sigma_{33} \end{pmatrix} \equiv \begin{pmatrix} \Sigma_{\phi\phi} & \Sigma_{\phi\psi} & \Sigma_{\phi\chi^1} \\ \Sigma_{\psi\phi} & \Sigma_{\psi\psi} & \Sigma_{\psi\chi^1} \\ \Sigma_{\chi^1\phi} & \Sigma_{\chi^1\psi} & \Sigma_{\chi^1\chi^1} \end{pmatrix} \tag{3}$$

Trivariate gaussian functions were fitted to the experimental data for each of the 20 amino acids (from all the proteins in Table I of the supplementary materials[32]). For each of the 12 clusters of data for each residue, the centroid and covariance matrices of the trivariate gaussian function associated with that cluster were computed. The centroid, $\boldsymbol{\mu}$, is equal to the mean of each dihedral angle. The centroids of each distribution were calculated using the procedure outlined in eqs. (4)–(7) [see discussion of eqs. (6) and (7) in Ref. 4] because the dihedral angle space is periodic, e.g., the mean of two angles $-179°$ and $179°$ is equal to $180°$, not zero. For each cluster, the mean was calculated by mapping the dihedral angle, x_n, to a point $\mathbf{y}_n$, where $\mathbf{y}_n$ is defined as:

$$\mathbf{y}_n = \begin{pmatrix} y_{1n} \\ y_{2n} \end{pmatrix} = \begin{pmatrix} \cos x_n \\ \sin x_n \end{pmatrix} \tag{4}$$

$$\cos \mu_i = \frac{1}{N} \sum_{n=1}^{N} \cos x_n, \qquad \sin \mu_i = \frac{1}{N} \sum_{n=1}^{N} \sin x_n \tag{5}$$

where N is the number of dihedral angles x_n in each set (i.e., ϕ, ψ, and χ^1) of a given type of residue, and μ_i represents the elements of the vector quantity $\mu = (\mu_\phi, \mu_\psi, \mu_{\chi^1})$.

$$\tan \mu_i = \frac{\sin \mu_i}{\cos \mu_i} = M \tag{6}$$

$$\mu_i = \tan^{-1} M \tag{7}$$

The values of Σ_{ij}, the elements of the covariance matrix of eq. (3), were estimated from the following equation[4]:

$$\Sigma_{ij} = \frac{1}{N-1} \sum_{k=1}^{N} \mathbf{pv}(x_{ik} - \mu_i)\mathbf{pv}(x_{jk} - \mu_j) \tag{8}$$

where μ_i and μ_j represent the elements of the vector quantity $\boldsymbol{\mu} = (\mu_\phi, \mu_\psi, \mu_{\chi^1})$.

The values of ϕ, ψ, and χ^1 each lie in the interval $-180° < x_i \leq 180°$. The form of the probability function does not take into account the fact that ϕ, ψ, and χ^1 all lie in the interval [180, −180] and that the shape of the ϕ, ψ, and χ^1 space is a hypertorus. To account for this, a function $\mathbf{pv}(x)$ is defined, where $\mathbf{pv}(x)$ returns the *principle value* of x. This is accomplished by adding or subtracting $360°$ to x until the final value lies between $180°$ and $-180°$. With the aid of eq. (8), the trivariate probability distribution function of eq. (2) may be rewritten as follows

$$f(\mathbf{x}) = \frac{1}{2\pi^{3/2}|\boldsymbol{\Sigma}|^{1/2}} \times \exp\left[-\frac{1}{2}\mathbf{pv}(\mathbf{x} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} \mathbf{pv}(\mathbf{x} - \boldsymbol{\mu})\right] \tag{9}$$

With the substitution of the function $\mathbf{pv}(\mathbf{x} - \boldsymbol{\mu})$ for $(\mathbf{x} - \boldsymbol{\mu})$, the gaussian function, $\mathbf{f}(\mathbf{x})$, is no longer properly normalized since the space has become that of a hypertorus. To use a one-dimensional example, the gaussian function $f(X)$ is defined for $X = +\infty$ to $X = -\infty$. But the space is now defined only from $\Theta = +180°$ to $\Theta = -180°$. To be mathematically proper, we should map the function, $f(X) \rightarrow g(\Theta)$, where $g(\Theta)$ has been normalized for a hypertoroidal space. But we have not performed this mapping operation which means that the tails of the gaussian function have been truncated and the function, $f(X)$, is no longer properly normalized. However, this introduces a small error since it is the tail of the gaussians that is affected. We are not interested in the actual dihedral angle probabilities but simply want a method that will generate suitable dihedral angles for starting structures for energy minimization; therefore, no correction for lack of normalization will be applied.

If the subscripts i and j are transposed in eq. (8), the expression is unchanged, which shows that the covaraince matrix, $\boldsymbol{\Sigma}$, must be symmetric; hence, the following relations hold

$$\begin{aligned} \Sigma_{12} &= \Sigma_{21} & \Sigma_{\phi\psi} &= \Sigma_{\psi\phi} \\ \Sigma_{23} &= \Sigma_{32} & \Sigma_{\psi\chi^1} &= \Sigma_{\chi^1\psi} \\ \Sigma_{13} &= \Sigma_{31} & \Sigma_{\phi\chi^1} &= \Sigma_{\chi^1\phi} \end{aligned} \tag{10}$$

The centroids and the covariance matrices for the experimental distributions of the ϕ, ψ, and χ^1 dihedral angles for each of the individual residues are shown in Table II of the Supplementary Material.[32] As an example, Table I shows the centroids and covaraince matrices for the Cys residue. The centroids and the covariance matrices for the cumulative distributions for all 20 amino acids are shown in Table II.

DESCRIPTION OF CONFORMATIONAL STATES

Backbone structure predictors typically define the backbone structural *states* of an amino acid residue as α-helical, β-sheet, or coil. In this article, we shall use the definition of backbone structure

TABLE II.
Distributions of Conformational States in the (ϕ, ψ, χ^1) Space for All Residues.[a]

N[b]	Conf.[c]	μ_ϕ (deg.)	μ_ψ (deg.)	μ_{χ^1} (deg.)	$\Sigma_{\phi\phi}$ (deg.)^2	$\Sigma_{\phi\psi}$ (deg.)^2	$\Sigma_{\phi\chi^1}$ (deg.)^2	$\Sigma_{\psi\psi}$ (deg.)^2	$\Sigma_{\psi\chi^1}$ (deg.)^2	$\Sigma_{\chi^1\chi^1}$ (deg.)^2
2216	αg^-	−74.41	−29.27	−68.30	357.42	−248.13	−36.82	402.00	25.10	229.07
711	αg^+	−82.61	−14.37	65.89	519.78	−272.41	13.28	368.76	−20.09	241.25
1186	αt	−65.42	−41.77	−176.05	243.71	−53.93	0.45	175.39	29.35	341.26
1847	ϵg^-	−103.62	137.15	−64.13	595.35	−14.97	−25.16	595.93	−4.62	214.91
732	ϵg^+	−121.64	157.76	63.11	936.52	−79.37	27.88	591.39	45.04	260.32
1461	ϵt	−104.47	126.06	−176.68	936.26	−62.19	−3.63	472.07	12.16	226.56
152	$\alpha^* g^-$	61.25	35.45	−63.35	190.88	−164.70	−23.30	396.02	72.75	291.26
9	$\alpha^* g^+$	56.64	39.93	52.87	1355.28	−752.54	204.86	1193.82	−436.28	851.10
44	$\alpha^* t$	65.42	41.58	−158.90	667.46	−95.99	−10.25	1093.95	184.00	393.09
14	$\epsilon^* g^-$	62.60	−172.91	−64.60	1936.71	−612.53	−197.49	2983.33	300.46	337.63
4	$\epsilon^* g^+$	54.05	−137.15	69.86	1670.52	856.37	153.13	1088.16	136.03	42.74
12	$\epsilon^* t$	95.62	−170.32	−159.30	1481.57	528.24	95.15	1848.25	−216.91	514.23

[a] The quantities μ_i and Σ_{ij} are defined in eqs. (7) and (8), respectively.
[b] N is the number of occurrences of the conformational state.
[c] The 12 conformational states are defined in the "Description of Conformational States" section.

that was adopted for use with the PRISM backbone structure predictor,[2-4] i.e., in terms of the backbone dihedral angles, ϕ and ψ, of each residue of the polypeptide. In this system, the backbone *conformational states* of the individual residues are classified as α, α^*, ϵ, and ϵ^*,[2-4] which correspond to rectangular regions of the ϕ–ψ map (see Fig. 1).[2-4]

Analysis of the dihedral angles, ϕ, ψ, and χ^1 showed that the data clustered primarily into 12 regions on the ϕ, ψ, and χ^1 space. Therefore, this space was partitioned into 12 regions. The influence of the backbone state on the side-chain state is well documented,[14, 16-19] and can be seen clearly by comparing the distributions for the αg^- state and the ϵg^- state from the cumulative frequency distribution for the ϕ, ψ, and χ^1 dihedral angles (Table II). Although the side-chain *conformational state* is g^- for both backbone states, there is a 4.17° difference in the mean value for χ^1.

With the inclusion of the side-chain dihedral angle, χ^1, the *conformational state* nomenclature must be expanded to include the side chain. Examination of the experimetnal data shows that the values of ϕ and ψ cluster in four regions of the ϕ, ψ space; that is why the ϕ, ψ space was divided into four regions, α, ϵ, α^*, ϵ^* (see Fig. 1).[2-4] The values of χ^1 cluster about $-60°, 60°, -180°$ (g^-, g^+, and t); accordingly, the χ^1 space is divided into three conformation states also named g^-, g^+, and t. Therefore, the ϕ, ψ, and χ^1 space was divided into 12 regions (viz., αg^-, αg^+, αt, $\alpha^* g^-$, $\alpha^* g^+$, $\alpha^* t$, ϵg^-, ϵg^+, ϵt, $\epsilon^* g^-$, $\epsilon^* g^+$, $\epsilon^* t$). The boundaries of the regions in the ϕ, ψ space are shown in Figure 1. The three side-chain *conformational states* encompass the following regions of the χ^1 space: $\chi^1 = -60° \pm 60°$ for the g^- region, $\chi^1 = 60° \pm 60°$ for the g^+ region, $\chi^1 = 180° \pm 60°$ for the t region.

SELECTION OF χ^1 CONFORMATIONAL STATES

In our initial test of the trivariate dihedral angle probability functions, we shall assume that the backbone *conformational states* are known, and must then select the *conformational states* for χ^1. Furthermore, in later applications of the trivariate distribution function to unknown protein structures, it will be necessary to predict the *conformational states* of both the backbone and χ^1. In either of these types of calculations, the *conformational states* of χ^1 must be identified. Therefore, we now consider the selection of the *conformational states* of χ^1.

A brute force approach would involve sampling the whole χ^1 space randomly. However, we introduce a more efficient procedure based on preferential sampling of regions of high probability by using a *prior* probability.

The prior probability is the probability of occurrence of a side-chain *conformational state given* that the backbone *conformational state* is known. By calculating the prior probabilities for each side-chain *conformational state*, high probability side-chain *conformational states* can be selected. The prior probability for residue type r, $p^r(i|j)$, where

p is the probability of occurrence of the ith side-chain conformation given that the backbone is in a *conformational state* j, can be calculated by the following equation

$$p^r(i|j) = \frac{f^r(i|j)}{f^r(g^+|j) + f^r(g^-|j) + f^r(t|j)} \quad (11)$$

$f^r(i|j)$ is the frequency of occurrence of the side-chain state, i ($i = g^-, g^+, t$), when the backbone is found in conformational state, j ($j = \alpha, \epsilon, \alpha^*, \epsilon^*$).

Let us suppose that a polypeptide sequence, S, has n residues, and that residue r has backbone conformational state α. There would be three possible side-chain conformational states, g^-, g^+, and t for χ^1 for this residue. For the n residues in the sequence there would be 3^n possible combinations of side-chain conformational states for the sequence. To explore the side-chain conformational space of this sequence adequately, on the order of 3^n starting conformations would have to be generated. A side-chain conformational state prediction which specifies the conformational state of every side chain for the sequences, S, would reduce the size of the conformation space that must be explored from on the order of 3^n to n.

Therefore, the side-chain comforational space was sampled by constructing high probability side-chain states. First, the *prior* probability for all the side-chain conformational states of the sequence, S, is calculated [see eq. (11)]. For example, the αg^-, αg^+, and αt conformational states occur with frequencies of 78, 16, and 26, respectively, for the Cys residue (see Table I). The prior probability $p^{cys}(g^-|\alpha) = 78/120$, $p^{cys}(g^+|\alpha) = 16/120$, and $p^{cys}(t|\alpha) = 26/120$. Listed in order of decreasing probability, the states are αg^-, αt, and αg^+. The probability of ocurrence, P_S, of a particular combination of side-chain conformational states for the amino acid sequence, S, can be calculated by obtaining the product of the prior probabilities, p_k^r, for each residue in the sequence.

$$P_S = \prod_{k=1}^{n} p_k^r(i|j) \quad (12)$$

The probabilities, P_S, can be calculated for all possible combinations of side-chain states for the entire polypeptide. Systematic exploration of the side-chain conformational space can be accomplished by choosing an appropriate number of the most probable conformational states. We have chosen the ten most probable side-chain states, i.e., the ten highest values of P_S. In this manner, high probability side-chain states can be sampled consistently. The most probable side-chain prediction associated with the native backbone conformational state for the *Streptococcal* fragment is listed in Table III. Only the most probable side-chain prediction for this sequence has been listed in the table, as an example, rather than showing the entire set of ten most probable conformational states for this protein. For comparison, the experimentally observed conformational states for χ^1 are also listed in Table III. The side-chain conformational states in the conformation of highest probability are predicted with 57% accuracy for protein G. The most probable side-chain prediction associated with the native backbone conformational state for *Antennapedia* is listed in Table IV. The side-chain conformational states for the *Antennapedia* fragment varied for the 20 different NMR structures; therefore, the experimental values are not listed in Table IV.

Once the backbone state and the side-chain conformational state have been chosen, the trivariate gaussian distribution associated with the given conformational state is used to generate values of ϕ, ψ, and χ^1 for subsequent energy minimization. Sampling is concentrated in regions of high probability, although it is still possible to sample *any* point in the ϕ, ψ, and χ^1 space.

PROBABILITY DISTRIBUTION FUNCTIONS

After the backbone and side-chain *conformational states* are identified, three different probability distribution functions are used to generate dihedral angles: the trivariate gaussian, bivariate gaussian, and uniform (or flat) probability distributions. The trivariate gaussian probability function generates ϕ, ψ, χ^1 dihedral angles and the bivariate gaussian function generates ϕ and ψ only. Both of these functions are biased toward experimental values, unlike the uniform (flat) probability distribution function. Given the region of the ϕ, ψ, and χ^1 space, the uniform function will generate values for these three dihedral angles uniformly over that region of conformational space. The uniform function always generates dihedral angles which are contained within the boundaries of the backbone and side-chain conformational state. Since the tails of the gaussian functions extend beyond the boundaries of a specified conformational state, it is still possible to sample *any* points in the ϕ, ψ, and χ^1 space using the trivariate or bivariate probability distribution functions.

TABLE III.
Listing of the Sequence, Native Backbone Conformational State, and the Most Probable Predicted Side-Chain Conformational States of χ^1 for the Binding Domain of *Streptococcal* Protein G.[a]

Residue	NB	S1	$(S1)_{exptl}$
MET1	α	g^-	g^-
THR2	ϵ	g^-	g^-
TYR3	ϵ	g^-	g^-
LYS4	ϵ	g^-	t
LEU5	ϵ	g^-	t
ILE6	ϵ	g^-	g^-
LEU7	ϵ	g^-	g^-
ASN8	α	g^-	g^-
GLY9	ϵ	g^-	t
LYS10	α	g^-	g^-
THR11	α	g^+	g^+
LEU12	ϵ	g^-	t
LYS13	ϵ	g^-	g^-
GLY14	ϵ	g^-	t
GLU15	ϵ	g^-	g^+
THR16	ϵ	g^-	t
THR17	ϵ	g^-	g^+
THR18	ϵ	g^-	t
GLU19	ϵ	g^-	t
ALA20	ϵ	g^-	t
VAL21	α	t	g^+
CYS22	ϵ	t	g^+
ALA23	α	g^-	t
ALA24	α	g^-	t
THR25	α	g^+	t
ALA26	α	g^-	t
GLU27	α	g^-	g^-
LYS28	α	g^-	g^-
VAL29	α	t	t
PHE30	α	t	g^-
LYS31	α	g^-	t
GLN32	α	g^-	t
TYR33	α	g^-	t
ALA34	α	g^-	t
ASN35	α	g^-	t
CYS36	α	g^-	g^-
ASN37	α	g^-	g^-
GLY38	α*	g^+	t
VAL39	ϵ	t	t
CYS40	ϵ	t	g^-
GLY41	α	t	t
GLU42	ϵ	t	g^-
TRP43	ϵ	g^-	g^-
THR44	ϵ	g^-	g^+
TYR45	ϵ	g^-	t
CYS46	ϵ	t	t
CYS47	α	g^-	t
ALA48	α	g^-	t
THR49	α	g^+	g^+
LYS50	α*	g^-	g^-
THR51	ϵ	g^-	g^-
PHE52	ϵ	g^-	g^-
THR53	ϵ	g^-	g^-
VAL54	ϵ	t	g^+
THR55	ϵ	g^-	g^-
GLU56	α	g^-	t

[a] The native backbone conformational state, labeled NB, for the binding domain of *Streptococcal* protein G, followed by the highest probability side-chain conformational state, labeled S1, associated with the backbone conformational state. α, ϵ, α*, and ϵ* are regions of the ϕ, ψ map and g^-, g^+, and t are the regions of the χ^1 dihedral angle space. $(S1)_{exptl}$ is the experimentally observed conformational states for χ^1 for the *averaged* structure.

The trivariate gaussian probability distribution functions developed here are extensions of the bivariate gaussian probability distribution functions in ϕ and ψ developed by Lambert and Scheraga,[2-4] and are very similar to the discrete probability grids developed by Evans *et al.*[5] Because the conformation of the backbone has been found to be correlated with the conformation of the side chain,[14, 16-19] the bivariate functions have been extended to include χ^1. The division of the ϕ, ψ, and χ^1 space into 12 regions, i.e., αg^-, αg^+, αt, etc., *automatically* includes the correlation found between the backbone and side-chain conformations.

Accurate representations of the distributions of the dihedral angles found in the PDB should use functions that include all the side-chain dihedral angles to build up statistically likely side-chain conformations including correlations with the conformations of the backbone and neighboring residues. However, the side-chain dihedral angles, χ^2, χ^3, χ^4, etc., are not included in the new probability distribution functions because of the paucity of data in the PDB for these dihedral angles. Instead, these dihedral angles are set to arbitrary values at the start of minimization to avoid the computational expense to search the whole conformational space.

Although continuous probability distribution functions for ϕ, ψ, and χ^1 increase the size of the conformational space, these functions are *more accurate* than discrete rotamer libraries. With some

TABLE IV.
Listing of the Sequence, Native Backbone Conformational States (NB), the Most Probable Predicted Side-Chain Conformational States of χ^1 (NS1) Associated with the Native Backbone, the Most Probable Predicted Backbone Conformational State (PB), and the Most Probable Predicted Side-Chain Conformational State (PS1) Associated with this Predicted Backbone Conformational State for the Mutant *Antenniapedia* (C39 → S) Homeodomain.[a]

Residue[b]	NB	NS1	PB	PS1
THR7	α	t	α	g^+
TYR8	ϵ	g^-	α	g^-
THR9	ϵ	g^-	α	g^+
ARG10	α	g^-	α	g^-
TYR11	α	g^-	α	g^-
GLN12	α	g^-	α	g^-
THR13	α	g^+	α	g^+
LEU14	α	g^-	α	g^-
GLU15	α	g^-	α	g^-
LEU16	α	g^-	α	g^-
GLU17	α	g^-	α	g^-
LYS18	α	g^-	α	g^-
GLU19	α	t	α	t
PHE20	α	t	α	t
HIS21	α	g^-	α	g^-
PHE22	α	t	α	t
ASN23	ϵ	t	α	g^-
ARG24	α	g^-	α	g^-
TYR25	ϵ	g^-	α	g^-
LEU26	ϵ	g^-	α	g^-
THR27	ϵ	g^-	α	g^+
ARG28	α	g^-	α	g^-
ARG29	α	g^-	α	g^-
ARG30	α	g^-	α	g^-
ARG31	α	g^-	α	g^-
ILE32	α	g^-	ϵ	g^-
GLU33	α	g^-	ϵ	g^-
ILE34	α	g^-	α	g^-
ALA35	α	g^-	α	g^-
HIS36	α	g^-	α	g^-
ALA37	α	g^-	α	g^-
LEU38	α	g^-	α	g^-
SER39	α*	g^-	α	g^+
LEU40	ϵ	g^-	α	g^-
THR41	ϵ	g^-	α	g^+
GLU42	α	t	α	t
ARG43	α	g^-	α	g^-
GLN44	α	g^-	α	g^-
ILE45	α	g^-	α	g^-
LYS46	α	g^-	α	g^-
ILE47	α	g^-	α	g^-
TRP48	α	g^-	α	g^-
PHE49	α	t	α	t
GLN50	α	g^-	α	g^-
ASN51	α	g^-	α	g^-
ARG52	α	g^-	α	g^-
ARG53	α	g^-	α	g^-
MET54	α	g^-	α	g^-
LYS55	α	g^-	α	g^-
TRP56	α	g^-	α	g^-
LYS57	α	g^-	α	g^-
LYS58	α	g^-	α	g^-
GLU59	α	g^-	α	g^-

[a] The native backbone conformational state (obtained from the experimental structures), labeled NB, for the mutant *Antenniapedia* (C39 → S) homeodomain is listed in the second column followed by the highest probability side-chain conformational state, labeled NS1, associated with the native backbone conformational state in the second column. The highest probability PRISM backbone conformational state prediction (PB) is listed in the fourth column and the highest probability side-chain conformational state for χ^1 (PS1) associated with the predicted backbone is listed in the fifth column. The states NS1 and PS1 were predicted with the aid of eqs. (11) and (12). α, ϵ, α*, and ϵ* are regions of the ϕ, ψ map and g^-, g^+, and t are the regions of the χ^1 dihedral angle space. The experimental values of the χ^1 conformational states have not been listed here because the conformational states for the 20 NMR structures varied.
[b] Residues 1–6 and 60–67 have not been listed in this table because all computations excluded these residues.

rotamer libraries, one idealized conformation is expected to represent the conformations which are found in a $\pm 20°$ range about the ideal conformation.[14, 18] Rotamer libraries often fail to take into consideration the distribution of conformations about the idealized rotamer conformation; therefore, low probability conformations which lie far from ideality may be accorded equal weight as more frequently occurring conformations. In one library, all values of χ^1 lying within $\pm 20°$ of 60° are considered to be in the g^+ state with *equal* probability while conformations which lie outside this range are assigned zero probability.[14] The continuous representation has been chosen over the discrete representation because it is inherently more accurate. Although the use of a continuous representation instead of a discrete one would *seem* to increase the size of the conformational search by increasing the size of the conformational space, in reality discrete representations encourage poor searches of the conformational space. For these reasons, the χ^1 side-chain dihedral angle is represented by a continuous probability distribution function rather than by a discrete function.

The current number of high resolution structures in the PDB sets a limit on the amount of data to be fit by empirical probability distributions. If more data were available, it would be possible to include more conformational correlations, e.g., between nearest neighbors of each residue. Because of the paucity of such data we have had to omit such possible correlations and therefore consider only the influence of the identity of a residue on its own backbone conformation. Thus, for the 17 residues with variable χ^1 values, and the remaining three residues (Gly, Ala, Pro), there are $17 \times 12 + 3 \times 4 = 216$ conformational states to be considered. For several of these conformational states, there are insufficient data to obtain reliable parameters. For example, the $\alpha^* g^-$ and $\alpha^* t$ conformational states for the Cys residue occur only four times and one time, respectively, and the $\alpha^* g^+$ state does not occur at all. It is necessary to decide if these "empty" regions are really empty or if they only appear empty because the data base is too small or biased. If the decision were made that the space is empty, there is the small possibility that a solution could be missed. Therefore, a cumulative probability distribution composed of statistics from *all* 20 amino acids was used for rare states (see Table II) rather than assume that these states never occur. This cumulative distribution overestimates the probability for this region of the conformational space but will not miss a solution. If there were fewer than nine occurrences of a particular conformational state, the cumulative distribution (Table II) was used instead of the distribution for the individual residues (Table I). But, even with the cumulative distribution, some conformational states are rare, e.g., the $\epsilon^* g^+$ conformational state which is found only four times in our data base. Since our test polypeptides were never found in the $\epsilon^* g^+$ state, we have not made any further approximations to fit a probability distribution function to this state.

Once all of the conformational states, αg^-, αg^+, αt, etc., have a properly fitted gaussian distribution, and the backbone and side-chain conformational states have been obtained, three-dimensional starting structures for minimization were generated.

GENERATION OF DIHEDRAL ANGLES

Once the *conformational state* has been specified, the probability distribution function associated with that conformational state was used to generate values for ϕ, ψ, and χ^1. Three different procedures were used to generate dihedral angles from the three different types of probability functions: uniform, bivariate, and trivariate.

Uniform Probability Distribution Functions

A linear congruential random number generator, subroutine RNUNF from IMSL,[35] will generate random numbers, u, uniformly in the interval $(0, 1)$. The random numbers which lie in the interval $(0, 1)$ are modified to lie in the interval of interest. As an example, consider the generation of a value for χ^1 with conformational state g^+. The g^+ region covers $60° \pm 60°$. The output of the random number generator is modified to generate values between 0° and 120° by letting $\chi^1 = 120u$. Values for dihedral angles other than ϕ, ψ, and χ^1 are assigned as discussed below.

Bivariate Probability Distribution Functions

The procedure to generate values of ϕ and ψ is described in Ref. 4. The procedure to use the bivariate gaussian probability distribution functions is very similar to that used with the trivariate probability distribution functions described below. The values of the side-chain dihedral angles (including χ^1) are listed in Table III of Ref. 4 and Table V of this study. The values of χ^1 were chosen so that the γ-atom of the side chain is *trans* to the main chain nitrogen atom. If there is branching at the C^β atom, then there are two conformations which are *trans* to the γ-atom and the choice was made arbitrarily.[4]

Trivariate Probability Distribution Functions

For trivariate gaussian distributions, generation of the values for the dihedral angles is a multistep process. First, three random numbers which lie uniformly on the interval $(0, 1)$ were generated. Let **u** designate the vector of these three numbers that is the output of the random number generator RNUNF, available from IMSL.[35] Second, the inversion technique[36, 37] was used to transform **u** to **z** where **z** has a unit trivariate gaussian distribution, i.e., one with a centroid equal to zero and covariance matrix equal to the identity matrix. The inversion technique[36, 37] is a standard method which enables any continuous distribution, $F(x)$, to be

TABLE V.
Initial Values of the Side-Chain Dihedral Angles.[a, b]

Residues	Side-chain conformation			
ALA[c]	$\chi^1 = 60.0°$			
CYS	$\chi^1 = 180.0°$			
ASP	$\chi^1 = 180.0°$			
GLU	$\chi^1 = 180.0°$			
PHE	$\chi^1 = 180.0°$	$\chi^2 = 90.0°$		
GLY[d]				
HIS	$\chi^1 = 180.0°$	$\chi^2 = 90.0°$		
ILE	$\chi^1 = -60.0°$	$\chi^{22} = 60.0°$	$\chi^{31} = 60.0°$	
LYS	$\chi^1 = 180.0°$	$\chi^5 = 60.0°$		
LEU	$\chi^1 = 180.0°$	$\chi^2 = 60.0°$	$\chi^{31} = 60.0°$	$\chi^{32} = 60.0°$
MET	$\chi^1 = 180.0°$	$\chi^4 = 60.0°$		
ASN	$\chi^1 = 180.0°$	$\chi^3 = 0.00°$		
PRO[e]	$\chi^1 = -68.8°$ or $-53.3°$			
GLN	$\chi^1 = 180.0°$	$\chi^4 = 0.00°$		
ARG	$\chi^1 = 180.0°$	$\chi^5 = 0.00°$	$\chi^{61} = 0.00°$	$\chi^{62} = 0.00°$
SER	$\chi^1 = 180.0°$			
THR	$\chi^1 = -60.0°$	$\chi^{22} = 60.0°$		
VAL	$\chi^1 = 180.0°$	$\chi^{21} = 60.0°$	$\chi^{22} = 60.0°$	
TRP	$\chi^1 = 180.0°$	$\chi^2 = 90.0°$		
TYR	$\chi^1 = 180.0°$	$\chi^2 = 90.0°$	$\chi^6 = 0.0°$	

[a] Except for the values listed here, all side-chain dihedral angles are initially set to 180° in all the starting conformations used in the energy minimization calculations. If the bivariate (ϕ and ψ only) probability distribution was used, the χ^1 dihedral angles from this table were used for the starting conformations; however, if *trivariate or uniform* (ϕ ψ, and χ^1) probability distributions were used to generate the dihedral angles for the starting conformations, then the χ^1 values listed here were not used.
[b] The values of the side-chain dihedral angles listed here are the same as those used by Lambert and Scheraga.[4] These arbitrary angles were chosen to avoid contact with the main chain.[4]
[c] The χ^1 dihedral angle for alanine was *always* set to this value regardless of the type of probability function used.
[d] Glycine does not have any side-chain dihedral angles.
[e] See the "Methods" section for a discussion of the procedure to generate dihedral angles for proline.

generated provided the inverse function $F^{-1}(u)$ can be calculated. $F^{-1}(u)$ is the inverse error function, $z_i = 2^{1/2} erf^{-1}(2u_i - 1)$.[36, 37] Finally, $\mathbf{z}$ was transformed to $\mathbf{x}$ which has a trivariate gaussian distribution with the centroid equal to $\boldsymbol{\mu}$ and covariance matrix equal to $\boldsymbol{\Sigma}$ as shown in eq. (13). We designate this final set of random numbers as $\mathbf{x}$, where $\mathbf{x}$ is the vector quantity for the set (ϕ, ψ, χ^1). Let $\mathbf{C}$ be the Cholesky factor of $\boldsymbol{\Sigma}$, where $\mathbf{C}\mathbf{C}^t = \boldsymbol{\Sigma}$ and $\mathbf{C}$ is lower triangular. Then,

$$\mathbf{x} = \mathbf{C}\mathbf{z} + \boldsymbol{\mu} \qquad (13)$$

Bivariate gaussian probability distribution functions were used to fit the dihedral angle distributions for Pro, Ala, and Gly. Except for minor modifications to reduce the dimensionality from three to two, the procedure to use the probability distribution functions for these three special amino acids is exactly the same as the procedure to use the trivariate gaussian probability distribution functions described above.

Assignment of Other Dihedral Angles

Until now, the discussion has focused on the generation of values for ϕ, ψ, and χ^1. However, *all* the dihedral angles must have values assigned before energy minimization. For both the uniform trivariate (ϕ, ψ, χ^1) and the trivariate gaussian probability distribution functions, the side-chain dihedral angles, χ^2, and beyond, were set to the arbitrary values shown in Table V. For the bivariate gaussian probability function, all the side-chain dihedral angle values, including χ^1, were selected from Table V. In addition, the trivariate gaussian probability distribution function was used to produce one set of starting structures in which the values of the side-chain dihedral angles, χ^2 and beyond, were set to 180°. This set of initial side-

chain values was used to evaluate the effect of the arbitrary side-chain values shown in Table V.

The side-chain dihedral angles shown in Table V are the same ones used by Lambert and Scheraga,[4] and were chosen to avoid contact with the main chain. For aromatic residues, χ^2 was set to 90° since this is the conformation commonly found in experimental structures.[4] The values of the backbone dihedral angle ω were initially set to 180° prior to energy minimization.

TEST SEQUENCES

Two proteins with neither disulfide bonds nor prosthetic groups, such as metal ions, were chosen to test our procedure. One structure is an α-class protein and the other is an α/β-class protein. High resolution NMR structures are available for both sequences. There is no crystal structure for this α-class protein, but a crystal structure is available for this α/β-class protein. However, the crystal structure differs from the NMR structure, primarily in the loop regions.[38, 39] Since, we are interested in the solution structure, the NMR structure will be used as the reference structure.

The first sequence is a mutant fragment of the *Antennapedia* (C39 → S) homeodomain (2HOA). A cysteine, residue 39 of the fragment, has been mutated to a serine. The wild-type *Antennapedia* homeodomain is a dimer, in which the monomers are joined by a disulfide bond involving the Cys-39 residues but, with the mutation of the cysteine, the structure is a monomer. The polypeptide is a fragment from a larger protein, but the fragment has been expressed from *E. coli.* and folds independently.[40-47] The *Antennapedia* mutant should be a suitable candidate for structure determination by conformational search. Although the fragment has 68 residues, the NMR data show structure only for residues 7 through 59. Since the NMR experiment fails to detect any structure for residues 0 through 6 and residues 60 through 67, the ends are probably free to move in solution.[47] Therefore, all calculations have been carried out only on residues 7 through 59. There were 20 solutions to the NMR distance constraints for the *Antennapedia* mutant[47]; these structures were then subjected to energy minimization with the AMBER potential[47, 48] to relieve steric overlaps. The final rms deviation for the backbone atoms (N, C^α, C') was 0.24 Å for the 20 final solutions.[47]

The second test sequence is also a fragment of a larger protein, the immunoglobulin binding domain of *Streptococcal* protein G (2GB1). This is a 56-residue fragment. The fragment has been expressed from *E. coli.* and folds independently. The three-dimensional structure has an α-helix packed against four strands of a β-sheet. It is noteworthy that this structure is extremely stable for its size.[38, 49-53] Sixty structures obtained by solution of the NMR distance constraints and simulated annealing are available. The rms deviation for the 60 structures was only 0.27 Å for the backbone atoms. Therefore, the *averaged* structure which was subjected to simulated annealing and hence is in a local minimum, and is included in the PDB, was used as the representative structure for this peptide.[50]

Examination of the structures obtained from the NMR experiments shows that the side-chain conformational states are quite variable; hence, we have not stated the percent accuracy of our side-chain conformational state predictions for the *Antennapedia* fragment. Unfortunately, small proteins are highly exposed to the solvent, and the problem will not disappear with the choice of another structure. Protein G appears to be an exception but only because the use of the *averaged* structure hides the variability of the side-chain conformations. Examination of the 60 individual solutions to the NMR distance constraints shows that the side-chain conformational states are variable. Numbers such as the percent accuracy of the side-chain conformational state prediction should be considered from a qualitative and not a quantitative point of view.

GENERATION OF *ANTENNAPEDIA* STARTING STRUCTURES

To generate starting conformations for energy minimization, it was assumed that the backbone conformational states (in terms of α, ϵ, α^*, and ϵ^*) for the *Antennapedia* (C39 → S) homeodomain were known with 100% accuracy. Once the backbone conformational states were available, predictions for the side-chain conformational states were made as described previously. The 10 most probable side-chain conformational states for the entire sequence were chosen. The backbone and predicted side-chain states combine to form ten backbone-side chain states which limit the conformational space of the starting structures. However, Table IV lists only the native backbone and the most probable side-chain conformational states for the *Antennapedia* fragment.

Four sets of 500 starting conformations each (a total of 2000), labeled "A"–"D", for energy minimization were generated for the mutant *Antenna-*

pedia (C39 → S) homeodomain. For each set of conformational states ("A"–"D"), 50 structures (in terms of dihedral angles) were generated from each of the 10 backbone and side-chain states for a total of 500 starting structures. Table IV shows the backbone conformational state and the most probable side-chain conformational state. The first set of starting structures, labeled "A," used the experimental backbone conformational states, the 10 most probable side-chain conformational states, the trivariate (ϕ, ψ, χ^1) dihedral angle probability distribution functions, and *all* the side-chain dihedral angles (except for χ^1) were set to 180° prior to energy minimization. The second set of starting structures, "B," used the same conformational states with the trivariate dihedral angle probability distribution functions, and the side-chain dihedral angle values (other than χ^1) were set to the values shown in Table V. These two sets of starting structures differed only in the initial values of the side-chain dihedral angles beyond χ^1. The third set of starting structures, "C," used the same backbone conformational states as "A" but with the *bivariate* dihedral angle probability distributions to generate starting values for ϕ and ψ. The side-chain dihedral angles, including χ^1, are shown in Table V. Starting conformations for set "D" were generated in an analogous manner to the other two dihedral angle probability distributions: the backbone and side-chain state, which specifies the region of the ϕ, ψ, and χ^1 map to be explored, was used in conjunction with the uniform (flat) dihedral angle probability distributions to generate values for ϕ, ψ, and χ^1 within the regions specified by the conformational state. The side-chain dihedral angles beyond χ^1 were set to values shown in Table V. The choice of conformational states, dihedral angle probability functions, and values chosen for all other dihedral angles for data sets "A"–"D" are summarized in Table VI.

GENERATION OF STARTING STRUCTURES FOR PROTEIN G

To test this procedure on a polypeptide with α/β structure, energy minimizations were carried out on the immunoglobulin binding domain of *Streptococcal* protein G. However, the full procedure to test the distributions was not carried out on the binding domain of *Streptococcal* protein G, since the results from *Antennapedia* showed that the trivariate gaussian probability function is best (see the Results section). Therefore, only the experimental backbone and the 10 most probable side-chain conformational states were used with the trivariate gaussian probability function to generate values of ϕ, ψ, and χ^1 (Table III shows the native backbone and most probable side-chain conformation). A total of 6000 starting conformations for energy minimization were generated. Except for χ^1, which was generated by the trivariate gaussian distribution, all other side-chain dihedral angles were set to the values shown in Table V.

ASSESSMENT OF STRUCTURE QUALITY

Energy-based searches of the conformational space usually produce a selection of low-energy

TABLE VI.
Conditions Used to Generate Sets of Starting Structures "A" Through "F" for Energy Minimization for the *Antennapedia* Homeodomain.

Data set	Backbone conformational state	Side-chain conformational state[b]	Method used to generate ϕ, ψ, χ^1	Side-chain dihedral angles beyond χ^1
A	Experimental	10 predicted	Trivariate gaussian probability function	Set to 180° except for χ^1
B	Experimental	10 predicted	Trivariate gaussian probability function	As listed in Table V except for χ^1
C	Experimental	10 predicted	Bivariate gaussian probability[c] function	As listed in Table V
D	Experimental	10 predicted	Trivariate uniform probability function	As listed in Table V except[d] for χ^1
E	10 predicted[a]	100 predicted	Trivariate gaussian probability function	As listed in Table V except for χ^1
F	10 predicted[a]	100 predicted	Bivariate gaussian probability[c] function	As listed in Table V

[a] The ten *most probable* PRISM predictions are always chosen.
[b] For data sets "A" through "D," the *same* 10 predicted side-chain *conformational states* are used with the experimental backbone conformational state. For *each* PRISM-predicted backbone conformational state with data sets "E" and "F," 10 side-chain conformational states were predicted for a total of 100 backbone / side-chain states in each set.
[c] The bivariate gaussian probability functions do not generate values for χ^1; however, Table V lists the values used for χ^1.
[d] χ^1 is sampled uniformly.

structures. Of these low-energy structures, those with relatively high rms deviations from the experimental structure are mixed in with structures with very low rms deviations. Distinguishing the lowest rms deviation structure on the basis of energetic considerations alone has always been difficult because of inaccuracies in the force field, and also the omission, in some cases, of the term for the free energy of hydration. However, some success has been achieved in distinguishing the structure of lowest rms deviation from a set of 39 low-energy ECEPP structures with estimates of the free energy of hydration based on the solvent-exposed surface area of the polypeptide conformation.[54] We use this solvent model to try to distinguish the structure of lowest rms deviation from the rest of the low-energy structures. Typically, the most native-like (lowest rmsd) solution is a low-energy structure but not necessarily the one of lowest energy. Arguments about the cause of this discrepancy usually point to the absence of hydration in the potential.[55] Studies, using BPTI, show that, by adding an empirical solvent free energy term to low-energy ECEPP/3 structures, it is possible to distinguish the most native-like structure from a selection of low-energy, higher rmsd structures.[54, 56] However, the BPTI structures in that study were very native-like with rms deviations of the heavy atoms from the 4PTI X-ray structure of 2.19 Å or less.[57] We apply an empirical solvent model, SRFOPT (**S**olvent sphere **R**adius **F**ixed, atomic solvation parameters **OPT**imized), a solvation model based on the solvent-exposed surface area of a conformation,[54] to the structures generated here, some of which are very far from native, to see if this solvent model could distinguish native-like from non-native-like low-energy structures. Solvent-exposed surface area models of hydration are based on the approximation that the solvent-exposed surface area of a molecule is a measure of the free energy of hydration, F_h.[54, 56-61]

$$F_h = \sum_{i=1}^{n} c_i A_i \qquad (14)$$

where A_i is the solvent-exposed surface area of the ith atom, $i = 1, \ldots, n$, and c_i is an empirically determined solvation surface free energy density. The SRFOPT set of parameters was chosen for this study because it was able to discriminate between the most native-like structure of BPTI from a selection of low-energy native-like structures.[54]

As another test, we apply the knowledge-based mean field potential[24-31] of Sippl and coworkers to the structures generated here to see if this method could successfully distinguish the native from non-native structures. The knowledge-based mean field method relies on the recognition that: (i) the set of distances between pairs of amino acids in a sequence contains sufficient information to define the three-dimensional structure of the sequence; and (ii) the experimental structures in the PDB, in the form of these sets of distances, can be used to estimate Boltzmann probabilities for polypeptide conformations.

The probability of occurrence of a particular state is given by eq. (15) where $d(s)$ is the set of distances between pairs of amino acids with a separation of i residues in a state or conformation, s, of the polypeptide. $P_i[d(s)]$ is the probability of finding any pair of amino acids separated by i residues at some distance, d. $E_i[d(s)]$ is the conformational energy and $\beta = 1/kT$, where k is the Boltzmann constant and T is the absolute temperature

$$P_i[d(s)] = \frac{\exp\{-\beta E_i[d(s)]\}}{Z_i[d(s)]} \qquad (15)$$

where $Z_i[d(s)] = \Sigma_s \exp\{-\beta E_i[d(s)]\}$. $P_i[d(s)]$ can be estimated from the data base of experimental structures. But the summation over all states s is unknown. By taking the *difference* between the state (conformation) of interest and a reference state, and assuming that the summation over both ensembles is similar, then the summation terms cancel and the difference in the energies can be calculated for a given conformation of a polypeptide. The difference in the energies or the pair potential, ΔE_i^{ab}, is the difference in energy between $E_i^{ab}[d(s)]$ and $E_i[d(s)]$, where $E_i^{ab}[d(s)]$ is the mean energy between a pair of amino acid types, a and b, where a and b refer to specific amino acid pairs, e.g., Ile-Met. a and b are separated by i residues in the sequence and are at a distance $d(s)$ from each other. E_i is the mean energy between *any* two amino acids separated by i residues at a distance $d(s)$

$$\begin{aligned}\Delta E_i^{ab} &= E_i^{ab}[d(s)] - E_i[d(s)] \\ &= -kT\{\ln(P_i^{ab}[d(s)]) - \ln(P_i[d(s)]) \\ &\quad + \ln(Z_i^{ab}[d(s)]) - \ln(Z_i[d(s)])\} \qquad (16)\end{aligned}$$

The reference state is the mean over all residues in the data base and does not differentiate between amino acid types.

The mean field potential is composed of two energy terms, the pair potential (discussed above) and a simplified solvent potential. The solvent-exposed surface area, which is related to the free energy of hydration, or an atom, i, can be estimated by the number of neighbors found within a sphere of radius, R. The number of neighbors is therefore a measure of the solvent-exposed surface area.[24-31] The frequencies of occurrence of neighbors around residues can be used to estimate a mean field solvation potential. The derivation of the solvation term is analogous to the derivation for the pair potential shown above.[24-31] The combined energy terms represent the mean field potential which gives a measure of the quality of a conformation.

The native structure of an amino acid sequence would be the structure with the lowest mean field energy. However, this would mean minimizing the mean field energy, a problem which is equivalent in computational magnitude to an energy-based search for the native structure. To circumvent this problem, the z-score, $z_{S,C}$, is constructed.

$$z_{S,C} = \frac{\left[E_T(S,C) - \bar{E}_T(S)\right]}{\sigma_S} \qquad (17)$$

where S is the peptide sequence; $E_T(S,C)$ is the mean field energy for the conformation, C; $\bar{E}_T(S) = \sum_C E_T(S,C)/(L - l + 1)$; L is the length of a "polyprotein"; l is the length of the peptide sequence, S; and σ_S is the standard deviation. The native structure should be the structure with the lowest energy $E_T(S,C)$, or the lowest z-score, $z_{S,C}$. An ensemble of conformations is required to calculate $\bar{E}_T(S)$. This ensemble is generated by a "polyprotein." The polyprotein is a construct created by joining three-dimensional peptide and protein structures from a data base of known structures using short linker polypeptides. Using only the backbone and the C^β atom, conformations that include features regularly found in globular proteins can be generated by sliding the sequence, S, along the length of the polyprotein. With the generation of this large ensemble of conformations, the relative quality of any conformation can be evaluated by comparison with the mean of energies from the ensemble. This method was able to distinguish the native structure of 41 of 65 small proteins from a large array of incorrectly folded structures.[25]

PREDICTION OF BACKBONE AND SIDE-CHAIN CONFORMATIONAL STATES

Until this point, we have assumed that the 100% correct backbone conformational state is available. This allows us to evaluate the different probability distribution functions for the side chains by comparing the final structures. We now consider the *prediction* of the backbone (and side-chain) conformational states and use PRISM[2-4] as a backbone prediction method.

The PRISM backbone structure prediction method is a pattern recognition scheme based on statistical data obtained from the PDB. The frequencies of occurrence of *tripeptide* backbone conformational states, such as $\alpha\alpha\alpha$, $\alpha\epsilon\alpha$, etc., are obtained from the PDB. Particular properties, such as the hydrophobicity or the α-helix propensity, of the tripeptides are mapped on to the property space, and probability distribution functions are developed for the distributions in the property space. Since these probability distributions are in the property space, any tripeptide conformational state with a suitable property profile can be considered in the prediction. This allows the PRISM scheme to compute probabilities and make multiple backbone conformational state predictions for any amino acid sequence.[2-4] The ability to generate multiple predictions is especially useful for energy-based searches of the conformational space. The most probable predicted backbone conformational state (82% accurate) along with the most probable side-chain conformational state for the *Antennapedia* fragment is listed in Table IV.

However, even if perfectly accurate backbone conformational states are assumed, the conformational state of the terminal amino acid residues presents a special problem. First, ϕ_i is defined by the position of C'_{i-1}, N_i, C_i^α, and C'_i.[62] However, for residues at the beginning of a chain, the C'_{i-1} atom may be replaced by an atom in an endgroup or it may not exist at all. A similar argument applies for ψ at the end of a chain. Second, terminal residues should have separate distribution functions to account for the differences in their environment in the sequence, but unfortunately the data set for terminal residues is sparse, and the inadequacy of the data set is further compounded

by experimental problems usually associated with terminal residues. For example, the ends of a chain may be free to move in solution, giving rise to especially poor resolution in the experimental data for the beginning/end residues. Because it would be difficult to construct a proper probability distribution for the beginning/end residues, the distributions developed for the middle residues are used.

Since the backbone structure predictor is not able to predict the conformational state of the first or the last residue in an amino acid sequence, it would be necessary to sample from all 12 conformational states associated with each of the first and the last residues. This procedure would greatly expand the conformational space to be searched. Additionally, the end residues are not always well defined in experimental or computational structure determinations. For these reasons, rather than expend a large amount of computational effort on these residues, we have arbitrarily chosen to set the backbone conformational state of the first and last residues to be α (see Tables III and IV).

GENERATION OF STARTING DIHEDRAL ANGLES USING PREDICTED BACKBONE STATES

The 10 most probable backbone conformational states generated from the PRISM backbone structure prediction scheme replace the native backbone conformational state. For each one of the 10 backbone predictions, the 10 most probable side-chain state predictions were obtained from the conformational state prediction scheme for χ^1, giving a total of 100 different backbone and side-chain conformational states. Each combination of the backbone and side-chain state predictions was used to generate five sets of starting dihedral angles for a total of 500 starting conformations for each set of dihedral angle distributions. As an example, Table IV shows the most probable backbone conformational state and the most probable side-chain conformational state associated with that backbone state for the *Antennapedia* sequence. Nine other PRISM-generated backbone state predictions and their associated side-chain states are used, but were not listed in Table IV to conserve space. The prediction accuracy ranges from 75% to 82% for the 10 backbone predictions with an average of 78% for the 10 backbone predictions. The average of 78% prediction accuracy is consistent with PRISM prediction accuracy for α-class proteins in general.[3] The set of structures, "E," for energy minimization, used the predicted backbone and predicted side-chain conformational states, of which Table IV is an example, with the trivariate gaussian dihedral angle probability distribution to generate values for ϕ, ψ, and χ^1. The values of the side-chain dihedral angles beyond χ^1 are shown in Table V. The set of structures, "F," used the same predicted backbone and side-chain conformational states as "E" but used the bivariate gaussian dihedral angle probability distributions to generate starting values for ϕ and ψ. The values of the side-chain dihedral angles from χ^1 onward are those listed in Table V. The values of the dihedral angles, ω, were all set to 180° initially. The choice of conformational states, probability functions, and set of dihedral angle values for those dihedral angles that were not generated from a probability function for data sets "E" and "F" are summarized in Table VI.

COMPUTATIONAL METHODS

All the structures generated for energy minimization were terminated with NH_2 at the amino end and with COOH at the carboxyl end of the sequence (ECEPP endgroups type 1 and 11, respectively).[20-23] The SUMSL (**S**ecant-type **U**nconstrained **M**inimization problem **SoL**ver)[63] routine was used. All of the backbone and side-chain dihedral angles were allowed to vary simultaneously during the course of the minimization.

All minimizations were carried out on the Kendall Square research computer (the KSR1), a highly parallel machine with 128 processors. The processors have a peak speed of 40 Mf. All minimizations were carried out with a version of the ECEPP/3 program which had been converted to take advantage of the parallel architecture of the KSR1.[64] A typical computation would minimize the energies of 16 starting conformations using four processors for the minimization of the energy of each conformation, for a total of 64 processors. A typical amount of cpu time required for a minimization of the energy of one conformation of the *Antennapedia* mutant is about 45 minutes if only one processor is used.

Solvent-exposed surface areas were computed with MSEED.[54, 65, 66] The probe radius was set to 1.4 Å, which is the radius of a water molecule. MSEED is a fast algorithm which calculates the solvent-exposed surface area.[66-68]

Results and Discussion

To assess the performance of the new dihedral angle probability distributions, we shall first compare the results obtained by using the experimental backbone states and predicted side-chain conformational states to generate starting values for energy minimization, using: (i) the trivariate probability distributions developed here for ϕ, ψ, and χ^1; (ii) the bivariate probability distribution developed by Lambert and Scheraga[4] for ϕ and ψ; as well as (iii) the uniform or flat distribution for ϕ, ψ, and χ^1. All three types of dihedral angle distributions were used to generate starting conformations suitable for minimization with the SUMSL minimizer[63] and the ECEPP/3 force field.[23] Assuming that the number of starting conformations is large enough, the quality of the final selection of low-energy minimized structures obtained and the lowest rmsd structure in particular should be an indication of the ability of each probability distribution to generate initial starting conformations that can lead to structures that are close to the global minimum.

ANTENNAPEDIA STRUCTURE

To compare the performance of the original bivariate dihedral angle distributions with the new trivariate distributions, starting conformations were generated for the mutant *Antennapedia* (C39 → S) homeodomain using each of the three (trivariate, bivariate, and uniform) dihedral angle distributions as described in the Methods section. Optimal superposition of two structures, and the rmsd of the main-chain heavy atoms (N, C^α, C') of each of 500 final structures with respect to each of the 20 experimental NMR structures was calculated using the method of Kabsch.[69, 70] The NMR structures are designated arbitrarily by the numbers 1 to 20, and the NMR structure that produces the lowest rmsd, i.e., is closest to one of the set of 500 calculated structures, is listed in column 2 of Table VII; the lowest *computed* rmsd structure with respect to the NMR structure listed in column 2 is listed in column 3 of Table VII. The ECEPP/3 energy of the *computed* structure with the lowest rmsd with respect to the NMR structure in column 2 is listed in column 4 of Table VII; its rank on the energy scale is given in column 5 of Table VII. The conformation of lowest energy is given in column 6 of Table VII; its rmsd with respect to the NMR structure in column 2, and its rank on this rmsd scale, are given in columns 7 and 8, respectively (Table VII). The various conditions used to generate the 500 initial conformations of each set of structures, "A"–"F," are listed in the footnotes of Table VII and summarized in Table VI.

The best computed class of rmsd structures were obtained by using the trivariate gaussian dihedral angle distributions; the bivariate gaussian dihedral angle distributions produced the second best class of rmsd structures. The unbiased or flat distribution produced the worst rmsd structures. The lowest rmsd structures obtained from the "A" and "B" sets of structures, the two sets of data which used the trivariate gaussian ϕ, ψ, and χ^1 dihedral angle distributions, are of good quality showing rms deviations from the NMR structures of 1.90 Å and 1.78 Å, respectively (see Fig. 2A and B). The two rmsd's obtained from "A" and "B" are too close in value to suggest which set of structures is superior in quality. The set of conformations, "C," which used the bivariate ϕ and ψ probability distributions, produced a lowest rmsd structure of 3.26 Å (see Fig. 2C). In comparison, the lowest rmsd structure produced by the uniform ϕ, ψ, and χ^1 distribution was 6.14 Å rmsd.

Table VII shows the rmsd and ECEPP/3 energy for the lowest rmsd structure and lowest energy structure obtained after minimization for the mutant *Antennapedia* (C39 → S) homeodomain. The procedure that found the lowest rmsd structures had explored the native area of the conformation space most efficiently. By this criterion, the lowest rmsd conformation was found by using the native backbone state and the 10 most probable side-chain states with the ϕ, ψ, and χ^1 trivariate gaussian dihedral angle probability functions.

However, examination of only the lowest rmsd structure from a set of data may be misleading. Insufficient exploration of the conformational space could still lead to an exceptionally low rmsd structure which might be a statistical artifact. Examination of the *distributions* of the rmsd's should be a more reliable indication of the quality of the computed conformations. Histograms of the 10,000 rms deviations computed for each set of 500 structures, "A"–"F," are shown in Figure 3. The mean and standard deviations for the distributions of the rms deviations for structures "A"–"D" are 9.95 ± 3.45 Å, 9.81 ± 3.37 Å, 10.56 ± 3.06 Å, and 10.57 ± 2.11 Å, respectively. The distribution of rmsd's

TABLE VII.
Results of the Energy Minimizations of the *Antennapedia* (C39 → S) Mutant.

Series[a]	NMR structure[b]	lowest rmsd conf.[c] (Å)	**E** of lowest rmsd conf.[d] (kcal / mol)	**E** rank of lowest rmsd conf.[e]	Lowest **E** conf.[f] (kcal / mol)	rmsd of lowest **E** conf.[g] (Å)	rmsd rank of lowest **E** conf.[h]
A[i]	18	1.90	−611.34	5	−615.28	5.33	48
B[j]	10	1.78	−583.92	44	−606.81	3.96	16
C[k]	18	3.26	−561.28	167	−608.76	5.13	22
D[l]	18	6.14	7.05×10^{24}	499	−493.04	9.34	191
E[m]	5	5.94	−612.08	64	−643.57	8.84	43
F[n]	18	5.46	−593.37	231	−649.33	9.34	56

[a] Each set of structures "A"–"F", used different conditions to generate starting conformations and contained a total of 500 starting conformations for each set (see Table VI).
[b] The NMR structure that produced the *lowest* rmsd with one of this set of 500 structures.
[c] The lowest rmsd of one of the 500 computed conformations compared to the NMR structure in column 2.
[d] Energy of the *computed* conformation with the lowest rmsd with respect to the NMR structure in column 2.
[e] Rank of the conformation whose energy is given in column 4.
[f] Computed energy of the lowest energy conformation.
[g] Compared to the NMR structure in column 2.
[h] Rank of lowest-energy conformation on the rmsd scale with respect to the NMR structure in column 2.
[i] This set of 500 minimizations used the native backbone chain state and the ϕ, ψ, and χ^1 probability distributions to generate the starting conformations. The side-chain dihedral angles, excluding χ^1, were *all* set to 180°. This set of structures was included primarily to show that the arbitrary choice of side-chain dihedral angles shown in Table V (and used in series B) does affect the final structures.
[j] This set of 500 minimizations used the native backbone chain state and the ϕ, ψ, and χ^1 probability distributions to generate the starting conformations. Since the χ^1 dihedral angles were generated by a probability distribution, the χ^1 dihedral angles listed in Table V were not used. All other side-chain dihedral angles were set to the values specified in Table V.
[k] This set of 500 minimizations used the native backbone chain state and the bivariate ϕ, ψ probability distributions to generate starting conformations. The initial side-chain dihedral angles, including the χ^1 dihedral angle, were set to the values specified in Table V.
[l] This set of structures was included primarily to serve as a baseline for the other sets of structures. The 500 minimized structures in this set used the native backbone chain state and a uniform distribution to generate starting conformations. The initial dihedral angles were generated randomly within the region of the ϕ, ψ, χ^1 space specified by the conformational state. The initial side-chain dihedral angles (except χ^1, which were generated by using the uniform distribution) are specified in Table V.
[m] This set of 500 minimizations used the 10 most probable PRISM backbone structure *predictions* and the ϕ, ψ, χ^1 probability distributions to generate starting conformations. Fifty starting conformations were generated from each of the 10 predictions. Side-chain dihedral angles beyond χ^1 were set to 180° except as noted in Table V. Since χ^1 is generated by a probability distribution, the values for χ^1 listed in Table V should be ignored.
[n] This set of 500 minimizations used the 10 most probable PRISM backbone *predictions* and the bivariate ϕ, ψ probability distributions to generate starting conformations. Fifty starting conformations were generated from each of the 10 predictions. All of the side-chain dihedral angles were set to 180° except as noted in Table V.

with the lowest means indicates the overall best exploration of the conformational space. But a large standard deviation may also indicate better exploration of the conformational space. For example, consider two distributions of rms deviations both with the same mean. If the first distribution has a much larger standard deviation than the second, then the first distribution contains a much larger number of both high and low rmsd structures than the second and is exploring a larger range of the conformational space. By these criteria, the dihedral angle probability distributions would be ranked in the following order (from best to worst) for the four sets of structures, "A"–"D": first, the trivariate gaussian probability distribution (sets "A" and "B"); second, the bivariate gaussian probability distribution (set "C"); and finally, the uniform trivariate probability distribution (set "D"). These are the same rankings obtained by using only the lowest rmsd structure (column 3 of Table VII) as an indication of quality. Evidently, 500 structures was a sufficiently large number of samplings to produce reliable results for the *Antennapedia* polypeptide. The mean and standard

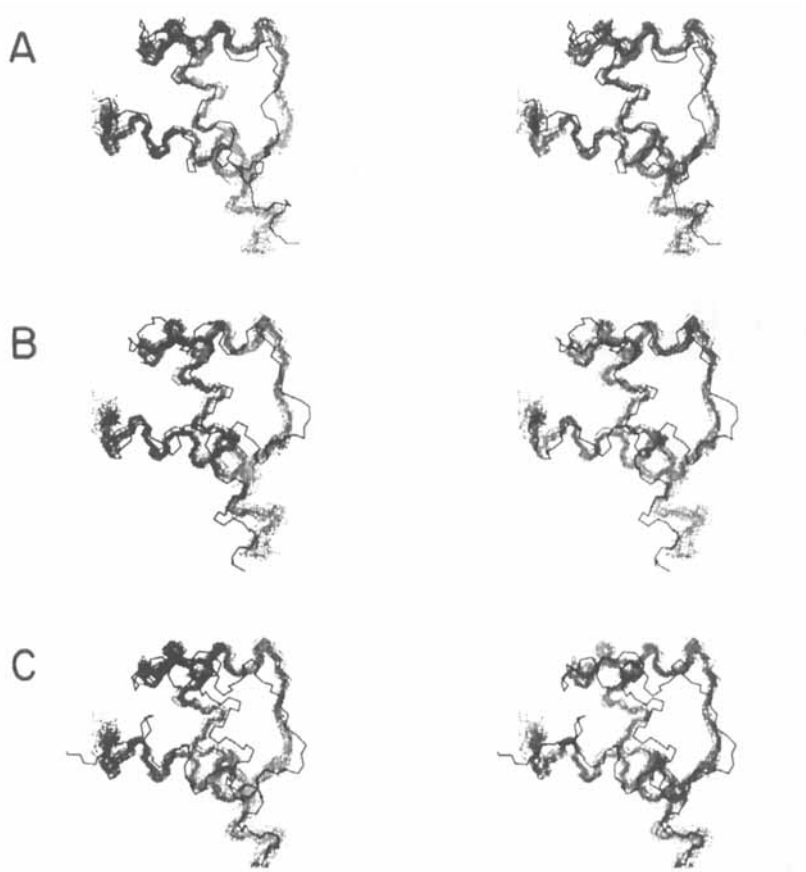

FIGURE 2. The lowest rmsd structures from the "A," "B," and "C" data sets. (A) The lowest rmsd structure from the "A" set of energy-minimized structures (solid line) superimposed on the 20 NMR structures[47] (dashed lines) of the *Antennapedia* mutant homeodomain. The lowest rmsd for this structure is 1.90 Å with respect to the 18th NMR structure. (B) The lowest rmsd structure from the "B" set of energy-minimized structures (solid line) superimposed on the 20 NMR structures[47] (dashed lines) of the *Antennapedia* mutant homeodomain. The lowest rmsd for this structure is 1.78 Å with respect to the 10th NMR structure. (C) The lowest rmsd structure from the "C" set of energy-minimized structures (solid line) superimposed on the 20 NMR structures[47] (dashed lines) of the *Antennapedia* mutant homeodomain. The lowest rmsd for this structure is 3.26 Å with respect to the 18th NMR structure.

deviation for "A," 9.95 ± 3.45 Å, and "B," 9.81 ± 3.37 Å are too close to declare one set of structures to be superior to the other.

A review of Table VII shows two sets of data, "A" and "B," where native-like structures were obtained. However, these two lowest rmsd structures are not those of lowest ECEPP/3 energy in their respective data sets. For the set of structures "A," there are four structures with energies lower than the lowest rmsd structure (see column 5 in Tables VII and VIII). For "B," there are 43 structures with lower energies than the lowest rmsd structures (see Table IX). This inability to distinguish native-like structures from other low-energy high-rmsd structures is a common problem with energy minimization schemes. We shall focus our attention on the structures with lower ECEPP/3 energies rather than on the lowest rmsd structure since it is the former structures which must be eliminated in our search for the native structure as the one of lowest ECEPP/3 energy.

Since the ECEPP/3 force field does not contain a term for the free energy of hydration, a simplified model of hydration is used to estimate this term. The solvent-exposed surface area[54, 66, 67] has been used to estimate the free energies of hydration for polypeptides and proteins. Vila *et al.* showed that free energies of hydration obtained from models, based on the solvent-exposed surface area, could be used to distinguish the most native-like structure of BPTI from a set of 39 low-energy structures obtained by minimization.[54]

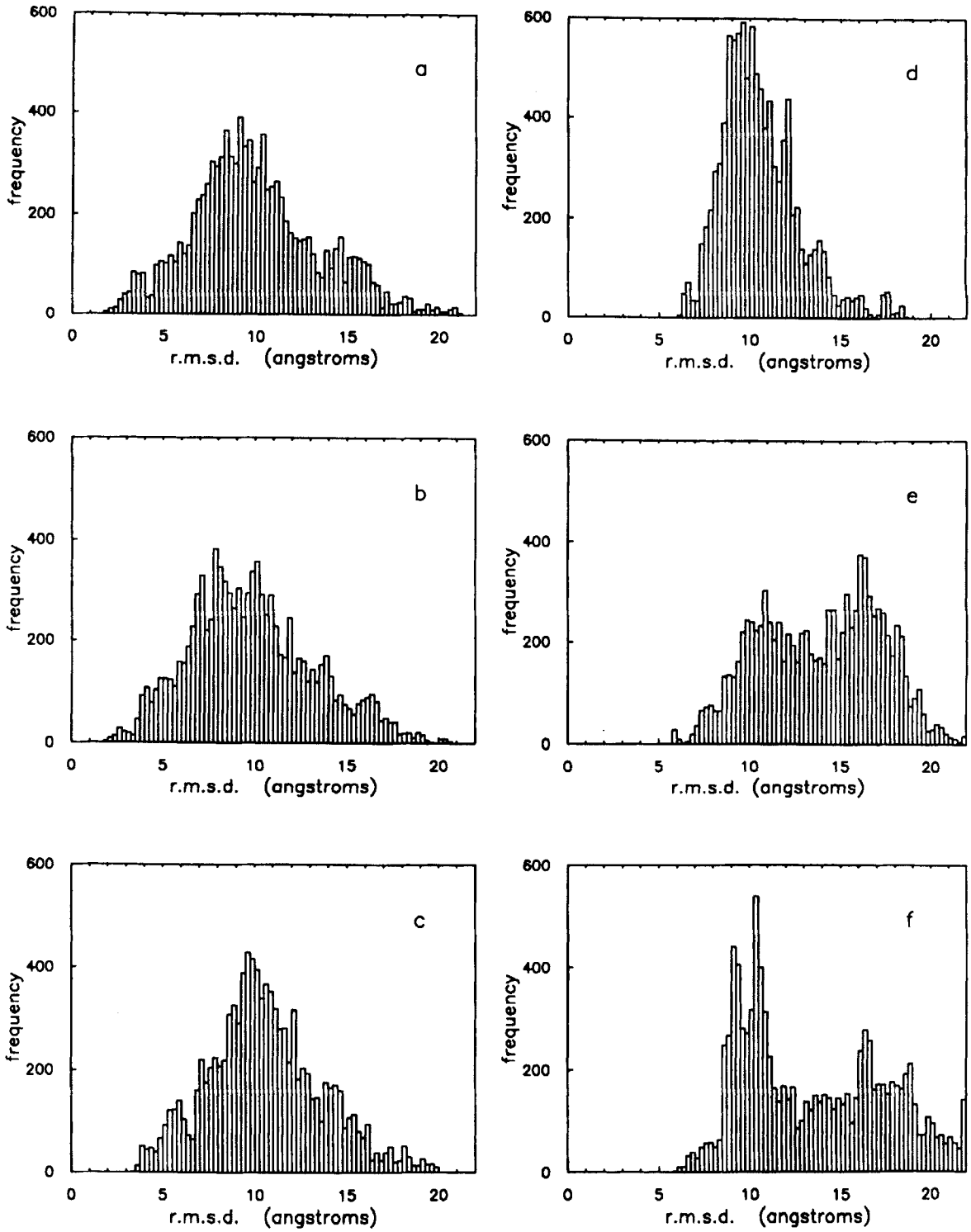

FIGURE 3. Histograms of the rms deviations for the sets of structures "A"–"F." The rms deviations for each of 500 energy-minimized structures with respect to all 20 NMR structures[47] (for a total of 10,000 calculations) are shown. (a) Histogram for the "A" set of structures. (b) Histogram for the "B" set of structures, etc.

TABLE VIII.
The Results of Adding the Free Energy of Hydration Term to the ECEPP/3 Energy for the Mutant *Antennapedia* (C39 → S) Homeodomain Using the SRFOPT Parameters.[a]

Conf.	E (kcal/mol)	E_{hyd} (kcal/mol)	rmsd[b] Å	Rank rmsd[b]
1	−615.28	−825.12	5.33	48
2	−612.79	−811.69	5.63	57
3	−611.81	−812.07	3.72	19
4	−611.47	−808.90	6.25	73
5	−611.34	−810.19	1.90	1

[a] The conformations are sorted in order of ECEPP/3 energies from highest to lowest, where **E** is the ECEPP/3 energy. E_{hyd} is the ECEPP/3 energy plus the hydration free energy obtained from the SRFOPT hydration model. The data presented in this table are derived from data set "A" in Table VII, which used the experimental backbone conformational state and the ϕ, ψ, and χ^1 gaussian probability distributions to generate the starting conformations for energy minimization. The starting side-chain dihedral angles beyond χ^1 were placed in "extended" positions with all side-chain dihedral angles set to 180°.

[b] These pertain to the calculations with **E**. The rms deviations were computed with respect to structure 18 of the 20 NMR structures.

Since the set of low-energy BPTI structures was of relatively high quality, we decided to test the SRFOPT model[54,57-61] to see if it could distinguish the lowest rmsd structure from a set of low-energy structures obtained by minimization with the ECEPP/3 force field. All of the energies of the minimized structures obtained by minimization with the ECEPP/3 force field were recalculated with the ECEPP/3 energies plus the free energy of hydration estimated with the SRFOPT model. Following the procedure of Vila *et al.*,[54] the structures were not reminimized with the free energy of hydration term. Tables VIII and IX show the effect of adding the SRFOPT hydration free energy to the lowest energy structures of data sets "A" and "B," respectively. It is clear from examination of E_{hyd}, the energy including the hydration, in Tables VIII and IX that, while the hydration term does affect the ranking of the minimized structures, the lowest rmsd structure has not become the lowest energy structure whereas it did in the study with BPTI.[54] It should be reiterated that the structures computed here were obtained by minimizing only the ECEPP/3 energy. Reminimization of the conformations using ECEPP/3 and SRFOPT together was not carried out; additional studies using a different solvation potential are being carried out but will not be reported here as this is beyond the scope of the present study.

Since the addition of an empirical hydration energy term did not improve the ordering of the conformations, we turned to the use of a knowledge-based potential.[24-31] *z*-Scores were computed for the same low-energy structures from data sets "A" and "B," whose ECEPP/3 energies and rmsd's are listed in Tables VIII and IX, respectively. Table X show the *z*-scores for the low-energy structures from data sets "A" and "B." The entries in Tables VIII–X were listed in order of ECEPP/3 energies, from the lowest to the highest. The lowest rmsd structure is always the last entry in these tables. The results from data set "A" are extremely encouraging (Table X, column A). The lowest rmsd structure (structure 5 of column A of Table X) is also the structure with the lowest *z*-score. However, for the *z*-scores with data set "B" (column B of Table X), the situation is less clear. Here, there are three structures (entry numbers 4, 20, and 30, respectively of column B of Table X) with lower *z*-scores than the lowest rmsd structure (entry 44 of column B of Table X). The lowest rmsd structure (entry 44 of column B) has a *z*-score of −4.02, whereas the three lowest *z*-scores are −4.45, −4.18, and −4.09 for entry numbers 4, 20, and 30, respectively, of column B. Further examination of the conformations numbered 4, 20, and 30 of column B reveals that, while these structures have relatively low rmsd's they are *not* the three that are the next lowest in rmsd; e.g., the entries labeled 5, 26, 29, 32, and 41 of column B all have lower rms deviations (see column 4 of Table IX). We have used both ECEPP/3 energies and *z*-scores, and neither criterion has identified the most native-like structure as the best structure. Comparison of the four lowest *z*-scores obtained from data set "B" with the *z*-scores obtained for the 20 NMR structures (see Table XI) shows that the lowest *z*-score from data set "B" (−4.45) cannot be distinguished from many of the *z*-scores obtained from NMR structures. Comparison of the ECEPP/3 energies for the lowest rmsd structures from sets "A" and "B" shows that the lowest rmsd structure from "A" also has a much lower energy (−611.34 kcal/mol) than any of the low-energy structures from "B" (lowest energy is −606.81 kcal/mol). The apparent failure of the knowledge-based mean field method to pick out the lowest rmsd structure from data set "B" suggests that the quality of the set of low-energy structures from "B" is too low

TABLE IX.

Results Obtained by Adding the Free Energy of Hydration Term to the ECEPP/3 Energy for the Mutant *Antennapedia* (C39 → S) Homeodomain Using the SRFOPT Parameters.[a]

Conf.	**E** (kcal/mol)	$\mathbf{E}_{hyd}$ (kcal/mol)	rmsd[b] (Å)	Rank rmsd[b]
1	−606.81	−803.92	3.96	16
2	−605.23	−784.13	3.81	17
3	−605.22	−811.76	9.56	257
4	−604.93	−792.52	3.48	7
5	−602.52	−785.47	3.01	4
6	−599.34	−793.67	6.91	98
7	−598.76	−792.05	6.35	70
8	−598.24	−805.59	9.99	287
9	−597.58	−786.56	11.13	347
10	−597.51	−805.79	9.87	278
11	−596.70	−806.25	3.81	12
12	−595.58	−787.60	7.97	163
13	−595.13	−816.58	5.77	55
14	−594.49	−792.00	6.24	69
15	−593.50	−790.13	4.97	34
16	−593.26	−778.08	6.41	73
17	−593.22	−797.19	6.39	72
18	−592.63	−786.44	8.13	174
19	−592.40	−803.14	4.75	28
20	−592.39	−791.71	5.34	45
21	−592.23	−812.10	12.44	393
22	−591.12	−768.01	6.95	102
23	−590.56	−791.40	10.64	321
24	−590.55	−786.83	8.00	164
25	−589.62	−771.88	8.65	206
26	−589.50	−791.71	3.88	14
27	−589.48	−777.57	5.70	54
28	−589.06	−781.46	10.32	308
29	−587.90	−792.49	2.59	2
30	−587.67	−786.42	4.64	26
31	−586.83	−773.76	5.60	50
32	−586.26	−785.69	2.61	3
33	−586.04	−790.35	6.63	82
34	−585.57	−810.11	12.83	405
35	−585.51	−778.43	9.16	235
36	−585.50	−801.37	5.81	57
37	−584.88	−778.10	9.43	251
38	−584.82	−792.56	9.59	260
39	−584.70	−791.66	10.03	290
40	−584.66	−791.44	5.11	37
41	−584.41	−779.16	3.18	5
42	−584.31	−788.53	4.22	20
43	−584.07	−795.98	7.44	129
44	−583.92	−782.96	1.78	1

[a] The conformations are sorted in order of ECEPP/3 energies from highest to lowest, where **E** is the ECEPP/3 energy. $\mathbf{E}_{hyd}$ is the ECEPP/3 energy plus the hydration free energy obtained from the SRFOPT hydration model. The data presented in this table are derived from data set "B" in Table VII, which used the experimental conformational state and the ϕ, ψ, and χ^1 gaussian probability distributions to generate the starting conformations for energy minimization. The starting side-chain dihedral angles beyond χ^1 were set to the values shown in Table V.

[b] These pertain to the calculations with **E**. The rms deviations were computed with respect to structure 10 of the 20 NMR structures.

TABLE X.

The *z*-Scores from the Mean Field Test of Sippl *et al.*[24-31] for the Mutant *Antennapedia* Homeodomain and the Binding Domain of Protein G.

Structure	*z*-Score *Antp*-A[a]	*Antp*-B[b]	Protein G[c]
1	−3.87	−3.00	−6.41
2	−3.78	−3.77	−6.06
3	−4.37	−3.14	−3.18
4	−4.03	−4.45	−3.36
5	−4.98	−3.96	−3.16
6		−3.42	−2.07
7		−2.83	−3.79
8		−2.44	−3.05
9		−2.64	−3.82
10		−3.09	−3.79
11		−3.45	−3.61
12		−3.61	−3.76
13		−3.23	−3.34
14		−3.11	−3.21
15		−2.17	−2.92
16		−3.12	−2.81
17		−3.98	−3.63
18		−3.12	−2.96
19		−3.91	
20		−4.18	
21		−2.40	
22		−2.83	
23		−3.04	
24		−2.91	
25		−3.18	
26		−3.98	
27		−3.07	
28		−2.81	
29		−3.85	
30		−4.09	
31		−3.85	
32		−3.98	
33		−2.84	
34		−2.34	
35		−1.93	
36		−3.46	
37		−3.13	
38		−2.21	
39		−2.08	
40		−2.77	
41		−3.18	
42		−3.20	
43		−2.32	
44		−4.02	

[a] Five conformations with the lowest ECEPP/3 energies from the "A" set of conformations for *Antennapedia* are listed in this column.

[b] The 44 structures with the lowest ECEPP/3 energy from the "B" set of conformations for *Antennapedia* are listed in this column.

[c] These are the 18 lowest energy structures of Protein G, listed in order of increasing energy. The 18th structure is the one with the lowest rmsd compared to the experimental structure.

for the the mean field to discern the most native-like structure. Although the two sets of low-energy structures from sets "A" and "B" would seem to be similar, both producing equally low rmsd structures, the final selection for best structure is the low rmsd (1.90 Å) structure from data set "A." The lowest rmsd structure from data set "A" has the lowest z-score of all the structures from both "A" and "B," and it has a lower ECEPP/3 energy than all the low-energy structures from "B." The starting structures of these two sets of structures, "A" and "B," differed only in the initial values of the side-chain dihedral angles beyond χ^1. For set "A," all of these dihedral angle values were set to 180°. For set "B," the values of these side-chain dihedral angles were taken from Table V. Although the values shown in Table V are arbitrary, they were chosen to avoid contact with the backbone atoms. However, the difference in the initial starting values for the dihedral angles beyond χ^1 appears to make little difference in the final structures.

We can estimate the ability of the three criteria (ECEPP/3 energies, ECEPP/3-plus-SRFOPT energies and the z-score) to identify low rmsd structures by using the correlation coefficient, r

$$r = \frac{(n\Sigma xy - \Sigma x \Sigma y)}{\left\{\left[n\Sigma x^2 - (\Sigma x)^2\right]\left[n\Sigma y^2 - (\Sigma y)^2\right]\right\}^{1/2}} \quad (18)$$

where n is the number of data points and x and y are the experimental quantities of interest. The correlation coefficients between the ECEPP/3 energies, the ECEPP/3-plus-SRFOPT energies and the z-scores for the 500 structures in each set with respect to the root-mean-square deviations from the 20 experimental structures for the *Antennapedia* protein are presented in Table XII. High energy structures (those structures with energies greater than zero) were removed since these structures are commonly high energy because of atomic overlaps and there is little correlation between energy and structure in these instances. Clearly, the knowledge-based mean field is the most highly correlated with the rmsd from the experimental structure, as indicated by the correlation coefficients of 0.691 and 0.656 for data sets "A" and "B," respectively.

TABLE XI.
The Results from the Mean Field Test of Sippl *et al.*[24-31] for the 20 NMR Structures of the Mutant *Antennapedia* Homeodomain.

Structure	z-Score	Structure	z-Score
1	−4.53	11	−4.97
2	−4.58	12	−4.72
3	−5.03	13	−5.24
4	−4.91	14	−5.14
5	−4.97	15	−4.50
6	−4.83	16	−5.10
7	−4.77	17	−4.74
8	−5.23	18	−5.09
9	−4.92	19	−5.00
10	−5.39	20	−4.93

PROTEIN G

The immunoglobulin binding domain of *Streptococcal* protein G is the second polypeptide used to test the new ϕ, ψ, and χ^1 probability distributions. This is an α/β class polypeptide and, as with the *Antennapedia* protein, the native backbone conformational state was used to try to recover the experimental structure by energy search using the ECEPP/3 force field. Six thousand starting structures were generated and their energies were minimized using the procedure outlined above. The experimental backbone states, and 10 most probable side-chain conformational states, were used to generate the starting dihedral angles for protein G. The experimental backbone and side-chain states and the most probable side-chain

TABLE XII.
The Correlation Coefficient for the rmsd vs. ECEPP/3 Energy, ECEPP/3 Energy with SRFOPT Hydration Term, and z-Scores Derived from the Mean Field Test for the Mutant *Antennapedia* (C39 → S) Homeodomain from the Sets of Structures "A" and "B."

	Correlation coefficient	
	Set "A"	Set "B"
ECEPP/3[a]	0.397	0.337
ECEPP/3 + SRFOPT[b]	0.054	0.047
z-score[c]	0.691	0.656

[a] From 311 and 320 conformations with ECEPP/3 energy less than zero for sets "A" and "B," respectively.
[b] From 311 and 319 conformations with ECEPP/3 plus SRFOPT energy less than zero for sets "A" and "B," respectively.
[c] From 311 and 320 conformations with ECEPP/3 energy less than zero for sets "A" and "B," respectively.

conformational states are shown in Table III. The resulting structure closest to the experimental one had an rmsd from the averaged NMR structure of 3.45 Å (see Table XIII and Fig. 4). The averaged NMR structure was used as the representative structure since Gronenborn *et al.* obtained 60 solutions and the averaged rmsd from the mean coordinates for the 60 solutions was 0.27 Å for all the backbone atoms.[50] As seen in Table XIII, the hydration term affects the ranking of the minimized structures, but the lowest rmsd structure has not become the lowest energy structure, as also observed for the *Antennapedia* protein.

When the knowledge-based mean field method[24-31] was applied to the low-energy solutions obtained from the energy minimizations with the ECEPP/3[23] force field for the binding domain of *Streptococcal* protein G, it was not able to determine the lowest rmsd structure (see Table X), i.e., the lowest rmsd structure was not the structure with the lowest *z*-score. As with other tests of the knowledge-based mean field method on structures obtained from X-ray crystallography,[29] this method occasionally failed to distinguish the native structure from other decoy structures. Most of these X-ray structures contained errors or were of low (2.8 Å) resolution.[29] Since the averaged NMR structure of protein G is the structure with the lowest calculated *z*-score (−7.53), the mean field method is able to discriminate between the experimental structure and the low-energy structures

TABLE XIII.
The Results of Adding the Free Energy of Hydration Term to the ECEPP/3 Energy for the Binding Domain of *Streptococcal* Protein G Using the SRFOPT Parameters.[a]

Conf.	**E** (kcal/mol)	$\mathbf{E}_{hyd}$ (kcal/mol)	rmsd[b] (Å)	Rank rmsd[b]
1	−400.81	−537.89	8.62	830
2	−391.06	−525.19	8.63	836
3	−390.25	−514.81	8.26	637
4	−388.24	−521.39	9.18	1160
5	−388.16	−503.63	7.58	370
6	−386.77	−503.27	8.44	751
7	−385.49	−515.36	6.38	106
8	−385.26	−505.63	9.98	1763
9	−383.71	−512.50	6.99	213
10	−383.44	−511.88	8.81	941
11	−383.42	−524.31	8.97	1024
12	−382.59	−513.66	5.05	11
13	−382.15	−514.88	7.03	220
14	−380.60	−514.95	6.27	90
15	−380.00	−510.10	7.56	362
16	−379.86	−509.79	7.33	295
17	−379.82	−515.40	12.77	3830
18	−379.29	−505.41	3.45	1

[a] The conformations are sorted in order of ECEPP/3 energies from highest to lowest, where **E** is the ECEPP/3 energy, $\mathbf{E}_{hyd}$ is the ECEPP/3 energy plus the hydration free energy obtained from the SRFOPT hydration model. The rmsd's are computed with respect to the *averaged* NMR structure.
[b] These pertain to the calculations with **E**.

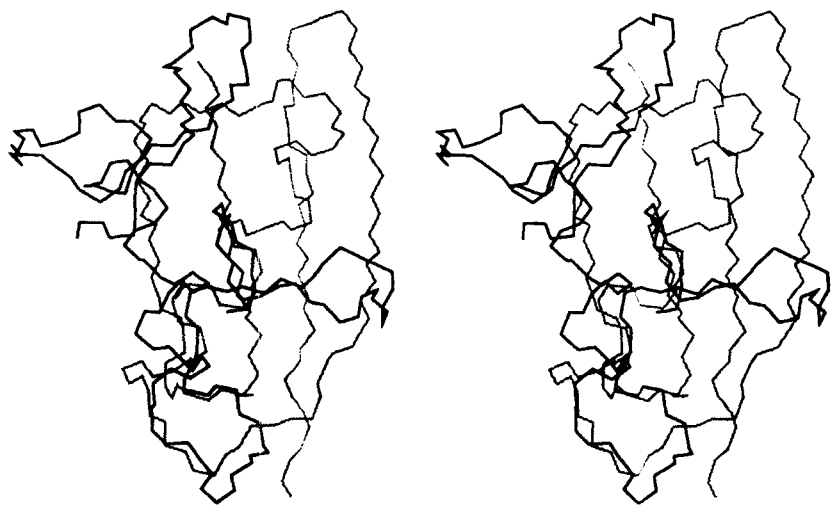

FIGURE 4. The lowest rmsd structure from the set of energy-minimized structures for the binding domain of *Streptococcal* Protein G (solid line) superimposed on the averaged NMR structure for the 60 solutions by Gronenborn *et al.*[50] (dashed line). The rmsd for this structure with respect to the averaged NMR structure is 3.45 Å.

obtained by energy minimization. The inability of this method to distinguish the lowest rmsd (3.45 Å) structure from the other low-energy structures means simply that the lowest rmsd energy-minimized structure is not *sufficiently* native-like for this method to detect any distinct difference from other low-energy structures. Since the mean-field potential is *knowledge based*, no information has been provided which would allow this method to distinguish between conformations that are far from native.

Why are the results for Protein G so poor? A total of 6000 starting structures were generated, and produced a low rmsd structure of 3.45 Å from the experimental structure. With the mutant *Antennapedia* fragment, only 500 structures were required to produce a low rmsd structure of 1.9 Å. Both polypeptides are similar in size. The difference lies in the backbone structure of the two protein fragments. α-Helices can be formed from the shorter range interactions of the amino acid sequence. Almost all of the information required to fold α-helical structures is contained in the backbone structure information. β-Sheets must form the correct pattern of hydrogen bonding between strands to assume the native structure correctly. In addition, the backbone conformational states and the trivariate gaussian probability functions used in this study do not distinguish between residues which form turns in the β-sheet and residues which are found in the strands of the β-sheet. Without this information, formation of the β-sheet becomes a random search through conformational space and much more computationally demanding.

RESULTS USING PREDICTED BACKBONE CONFORMATIONAL STATES

Using the PRISM predicted backbone conformational states, the lowest rmsd structures obtained for the *Antennapedia* (C39 → S) mutant homeodomain after energy minimization was 5.94 Å and 5.46 Å for the "E" and "F" sets of structures, respectively (see data sets "E" and "F" in Table VII and Fig. 5). The trivariate gaussian probability function was used to generate the ϕ, ψ, and χ^1 dihedral angles for the sets of structures, "E" and "B." Comparison of the histograms of "E" with those of "B" (see Fig. 3) shows a large difference in the distributions obtained. Similarly, the bivariate gaussian probability function was used to generate starting structures for the sets "F" and "C." The large rmsd's obtained from "E" and "F" (5.94 Å and 5.46 Å, respectively) show the effect of the inaccuracies in the backbone conformational state prediction.

Comparison of the lowest rmsd's obtained for the sets of structures "E" and "F" (5.94 Å and 5.46 Å, respectively) with the *mean* of the rmsd's 13.56 ± 3.95 Å for "E" and 14.20 ± 3.29 Å for "F" (see Fig. 3 for the histograms of the rmsd's for data sets "E" and "F"), reveals conflicting results. The trivariate gaussian distribution used to generate the initial dihedral angles for "E" has the lower mean rmsd, but the lowest rmsd structure is found in set "F." The lower rmsd structure for the bivariate probability distribution (set "F") may be a statistical artifact.

It is clear from these results that the backbone structure prediction procedure must become much more accurate before the improvement in the dihedral angle distributions can be used to predict the native structure of a polypeptide. For both the bivariate and trivariate dihedral angle probability distributions, the inaccuracies in the conformational state prediction defeats any increase in the accuracy of the distributions since sampling will now be concentrated in the incorrect portions of the conformational space. In fact, the low accuracy in the backbone conformational state prediction (approximately 78%), combined with the bivariate ϕ, ψ distributions may lead to a few starting structures that are slightly better than those obtained with the trivariate probability distribution functions, because the trivariate probability distribution functions produce starting values that are more highly concentrated in selected regions of the conformational space. This could explain why the lowest rmsd structure from the set of structures "E" had a higher rmsd than the lowest rmsd structure from "F." As a result, energy searches using the trivariate gaussian probability distributions explore *less* conformational space than the bivariate ϕ, ψ dihedral angle probability distributions. With highly accurate conformational states, the trivariate gaussian probability distributions can find lower rmsd structures than the bivariate gaussian probability distributions, but with inaccurate conformational states, the trivariate gaussian probability distributions search less conformational space and are therefore less likely to overcome mistakes in the conformational state information.

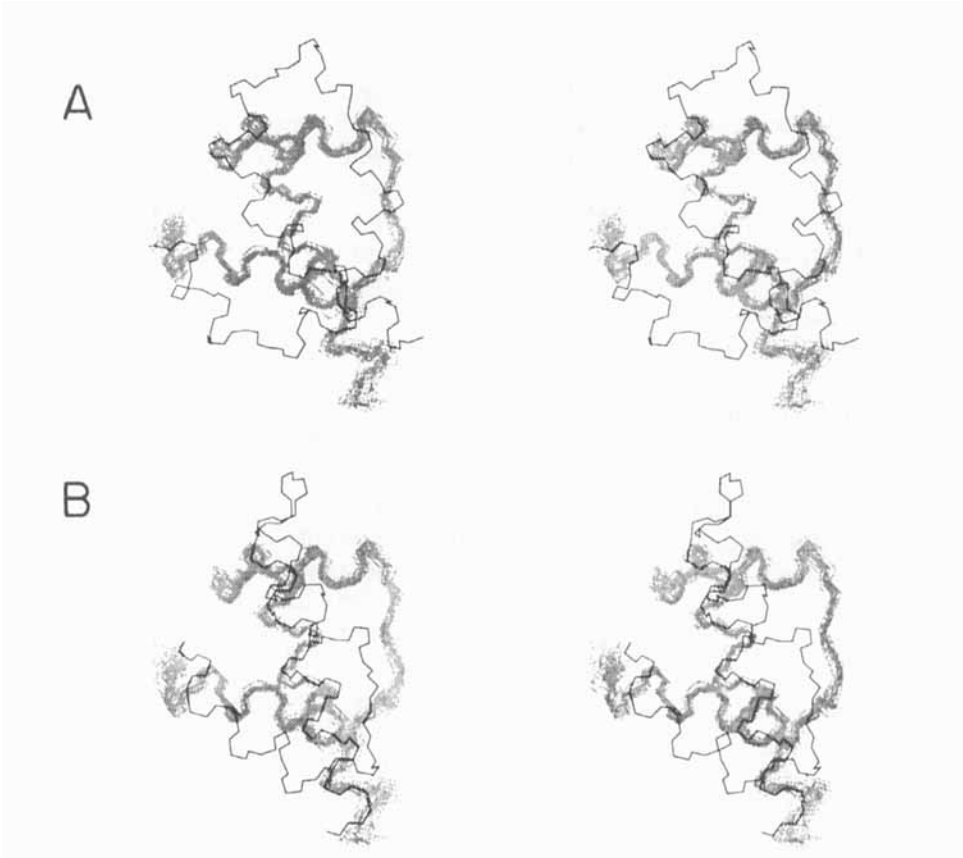

FIGURE 5. The lowest rmsd structures from the "E" and "F" sets of energy-minimized structures. (A) The lowest rmsd structure from the "E" set of energy-minimized structures (solid line) superimposed on the 20 NMR structures[47] (dashed lines) of the *Antennapedia* mutant homeodomain. The lowest rmsd for this structure is 5.94 Å with respect to the fifth NMR structure. (B) The lowest rmsd structure from the "F" set of energy-minimized structures (solid line) superimposed on the 20 NMR structures[47] (dashed lines) of the *Antennapedia* mutant homeodomain. The lowest rmsd for this structure is 5.46 Å with respect to the 18th NMR structure.

Conclusions

We have shown that, given an *accurate* backbone *conformational state* prediction, it is possible to find a native-like structure for α-helical proteins using the trivariate gaussian dihedral angle probability distributions. Current backbone structure prediction accuracies are too low for the dihedral angle probability distribution functions to be used to predict native structures from the amino acid sequence, which is the ultimate goal of protein folding studies. However, recent results of Srinivasan and Rose[71] suggest that the quality of backbone predictions can be greatly improved. Unfortunately, for proteins with β-sheet structure, even completely accurate knowledge of the backbone structure leads to structures with poor quality. For β-sheet structures, more information would be required for a successful structure prediction. At the very least, the turns in the β-sheet must be identified and separate dihedral angle probabilities should be developed for this class of conformations.

However, the dihedral angle probability distribution functions developed here will still be useful even without a 100% backbone structure prediction. If the length of the unknown sequence is short, then the trivariate gaussian distribution can be used to generate initial values of the dihedral angles ϕ, ψ, and χ^1 for all 12 states per residue (for subsequent energy minimization) in an exploration of the conformational space.

We have shown that it is possible to recover a native-like structure by an energy search, and that it is possible to discern a native-like structure from a host of equally low-energy structures. Unfortunately, the ability to select the native structure depends on the quality of the set of final low-en-

structure with a 1.78-Å rmsd was not clearly identified as a native-like structure by any of the three test criteria (ECEPP/3, ECEPP/3 and SRFOPT, and z-score). We now require a method that will distinguish between an array of poor quality albeit somewhat native-like structures, an obviously more stringent performance criterion than distinguishing highly resolved structures from poorly resolved structures.

Acknowledgments

We thank M. H. Lambert and D. M. Rothwarf for helpful discussions; S. Talluri for assisting with the selection of high resolution NMR and X-ray structures for inclusion in our data base; J. Kostrowicki for discussions pertaining to the trivariate probability function; M. J. Sippl for sending us the programs and materials for the mean-field method; D. R. Ripoll who converted the ECEPP program to use the parallel capacity of the KSR computer; G. M. Clore and K. Wüthrich and their coworkers for sending us the coordinates of the binding domain of protein G and the mutant *Antennapedia* homeodomain, respectively, before the structures became available from the PDB.

This work was supported at Cornell University by research grants from the National Institute of General Medical Sciences of the National Institutes of Health (GM-14312) and from the National Science Foundation (DMB90-15815). The computations were carried out on the KSR parallel computer at the Cornell National Supercomputer Facility, a resource of the Cornell Center for Theory and Simulation in Science and Engineering, which receives major funding from the National Science Foundation and IBM Corporation, with additional support from New York State and members of its Corporate Research Institute. The KSR facility was funded in part by the National Institutes of Health.

Betty Cheng was supported by an NIH Biotechnology traineeship from January 1991 through February 1994.

References

1. M. Vásquez, G. Némethy, and H. A. Scheraga, *Chem. Rev.*, **94**, 2183–2239 (1994).
2. M. H. Lambert and H. A. Scheraga, *J. Comput. Chem.*, **10**, 770–797 (1989).
3. M. H. Lambert and H. A. Scheraga, *J. Comput. Chem.*, **10**, 798–816 (1989).
4. M. H. Lambert and H. A. Scheraga, *J. Comput. Chem.*, **10**, 817–831 (1989).
5. J. S. Evans, A. M. Mathiowetz, S. I. Chan, and W. A. Goddard III, *Prot. Sci.*, **4**, 1203–1216 (1995).
6. A. M. Mathiowetz and W. A. Goddard III, *Prot. Sci.*, **4**, 1217–1232 (1995).
7. F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, *J. Mol. Biol.*, **112**, 535–542 (1977).
8. J. Janin, S. Wodak, M. Levitt, and B. Maigret, *J. Mol. Biol.*, **125**, 357–386 (1978).
9. M. N. G. James and A. R. Sielecki, *J. Mol. Biol.*, **163**, 299–361 (1983).
10. J. W. Ponder and F. M. Richards, *J. Mol. Biol.*, **193**, 775–791 (1987).
11. N. L. Summers, W. D. Carlson, and M. Karplus, *J. Mol. Biol.*, **196**, 175–198 (1987).
12. N. L. Summers and M. Karplus, *J. Mol. Biol.*, **210**, 785–811 (1989).
13. C. Lee and S. Subbiah, *J. Mol. Biol.*, **217**, 373–388 (1991).
14. H. Schrauber, F. Eisenhaber, and P. Argos, *J. Mol. Biol.*, **230**, 592–612 (1993).
15. F. Eisenmenger, P. Argos, and R. Abagyan, *J. Mol. Biol.*, **231**, 849–860 (1993).
16. R. L. Dunbrack, Jr. and M. Karplus, *J. Mol. Biol.*, **230**, 543–574 (1993); *Struct. Biol.*, **1**, 334–340 (1994).
17. R. Tanimura, A. Kidera, and H. Nakamura, *Prot. Sci.*, **3**, 2358–2365 (1994).
18. A. Nayeem and H. A. Scheraga, *J. Prot. Chem.*, **13**, 283–296 (1994).
19. M. Vásquez, *Biopolymers*, **36**, 53–70 (1995).
20. F. A. Momany, R. F. McGuire, A. W. Burgess, and H. A. Scheraga, *J. Phys. Chem.*, **79**, 2361–2381 (1975).
21. G. Némethy, M. S. Pottle, and H. A. Scheraga, *J. Phys. Chem.*, **87**, 1883–1887 (1983).
22. M. J. Sippl, G. Némethy, and H. A. Scheraga, *J. Phys. Chem.*, **88**, 6231–6233 (1984).
23. G. Némethy, K. D. Gibson, K. A. Palmer, C. N. Yoon, G. Paterlini, A. Zagari, S. Rumsey, and H. A. Scheraga, *J. Phys. Chem.*, **96**, 6472–6484 (1992).
24. M. J. Sippl, *J. Mol. Biol.*, **213**, 859–883 (1990).
25. M. Hendlich, P. Lackner, S. Weitchus, H. Floeckner, R. Froschauer, K. Gottsbacher, G. Casari, and M. J. Sippl, *J. Mol. Biol.*, **216**, 167–180 (1990).
26. M. J. Sippl, M. Hendlich, and P. Lackner, *Prot. Sci.*, **1**, 625–640 (1992).
27. G. Casari and M. J. Sippl, *J. Mol. Biol.*, **224**, 725–732 (1992).
28. M. J. Sippl, S. Weitckus, *Prot. Struct. Funct. Genet.*, **13**, 258–271 (1992).
29. M. J. Sippl, *Prot. Struct. Funct. Genet.*, **17**, 355–362 (1993).
30. M. J. Sippl, *J. Comp.-Aided Mol. Des.*, **7**, 473–501 (1993).
31. M. J. Sippl, M. Jaritz, M. Hendlich, M. Ortner, and P. Lackner, *Statistical Mechanics, Protein Structure, and Protein Substrate Interactions*, S. Doniach, Ed., Plenum Press, New York, 1994, pp. 297–315.

32. Supplementary Material is available from the authors upon request or via the Internet (see footnote * on page 1453).
33. S. B. Needleman and C. D. Wunsch, *J. Mol. Biol.*, **48**, 443–453 (1970).
34. T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, 2nd ed., Wiley, New York, 1984, p. 21.
35. IMSL, *IMSL Users Manual, Math/Library, Vol. 3*, Houston, TX, 1987, pp. 1111–1112.
36. P. Bratley, B. L. Fox, and L. E. Schrage, *A Guide to Simulation*, 2nd ed., Springer, New York, 1987, pp. 147, 161–162.
37. D. E. Knuth, *The Art of Computer Programming, Vol. II: Seminumerical Algorithms*, 2nd ed., Addison-Wesley, Reading, MA, 1981, pp. 116–117.
38. T. Gallagher, P. Alexander, P. Bryan, and G. L. Gilliland, *Biochemistry*, **33**, 4721–4729 (1994).
39. K. Kato, L.-Y. Lian, I. L. Barsukov, J. P. Derrick, H. Kim, R. Tanaka, A. Yoshino, M. Shiraishi, I. Shimada, Y. Arata, and G. C. K. Roberts, *Structure*, **3**, 79–85 (1995).
40. J. C. W. Shepherd, W. McGinnis, A. E. Carrasco, E. M. De Robertis, and W. J. Gehring, *Nature*, **310**, 70–71 (1984).
41. M. Müller, M. Affolter, W. Leupin, G. Otting, K. Wüthrich, and W. J. Gehring, *EMBO J.*, **7**, 4299–4304 (1988).
42. G. Otting, Y. Q. Qian, M. Müller, M. Affolter, W. Gehring, and K. Wüthrich, *EMBO J.*, **7**, 4305–4309 (1988).
43. Y. Q. Qian, M. Billeter, G. Otting, M. Müller, W. J. Gehring, and K. Wüthrich, *Cell*, **59**, 573–580 (1989).
44. M. P. Scott, J. W. Tamkun, and G. W. Hartzell III, *Biochim. Biophys. Acta*, **989**, 25–48 (1989).
45. M. Affolter, A. Percival-Smith, M. Müller, W. Leupin, and W. J. Gehring, *Proc. Natl. Acad. Sci. USA*, **87**, 4093–4097 (1990).
46. M. Billeter, Y. Q. Qian, G. Otting, M. Müller, W. J. Gehring, and K. Wüthrich, *J. Mol. Biol.*, **214**, 183–197 (1990).
47. P. Güntert, Y. Q. Qian, G. Otting, M. Müller, W. Gehring, K. Wüthrich, *J. Mol. Biol.*, **217**, 531–540 (1991).
48. P. K. Weiner and P. A. Kollman, *J. Comp. Chem.*, **2**, 287–303 (1981); S. J. Weiner, P. A. Kollman, D. T. Nguyen, and D. A. Case, *J. Comput. Chem.*, **7**, 230–252 (1986).
49. J. M. Sturtevant, *Annu. Rev. Phys. Chem.*, **38**, 463–488 (1987).
50. A. M. Gronenborn, D. R. Filpula, N. Z. Essig, A. Achari, M. Whitlow, P. T. Wingfield, and G. M. Clore, *Science*, **253**, 657–661 (1991).
51. G. M. Clore and A. M. Gronenborn, *J. Mol. Biol.*, **223**, 853–856 (1992).
52. B. Leiting, R. De Francesco, L. Tomei, R. Cortese, G. Otting, and K. Wüthrich, *EMBO J.*, **12**, 1797–1803 (1993).
53. T. A. Ceska, M. Lamers, P. Monaci, A. Nicosia, R. Cortese, and D. Suck, *EMBO J.*, **12**, 1805–1810 (1993).
54. J. Vila, R. L. Williams, M. Vasquez, and H. A. Scheraga, *Prot. Struct. Funct. Genet.*, **10**, 199–218 (1991).
55. D. H. Kitson, F. Avbelj, J. Moult, D. T. Nguyen, J. E. Mertz, D. Hadzi, and A. T. Hagler, *Proc. Natl. Acad. Sci., USA*, **90**, 8920–8924 (1993).
56. R. L. Williams, J. Vila, G. Perrot, and H. A. Scheraga, *Prot. Struct. Funct. Genet.*, **14**, 110–119 (1992).
57. D. R. Ripoll, L. Piela, M. Vásquez, and H. A. Scheraga, *Prot. Struct. Funct. Genet.*, **10**, 188–198 (1991).
58. Y. K. Kang, G. Némethy, and H. A. Scheraga, *J. Phys. Chem.*, **91**, 4105–4109 (1987).
59. Y. K. Kang, G. Némethy, and H. A. Scheraga, *J. Phys. Chem.*, **91**, 4109–4117 (1987).
60. Y. K. Kang, G. Némethy, and H. A. Scheraga, *J. Phys. Chem.*, **91**, 4118–4120 (1987).
61. Y. K. Kang, K. D. Gibson, G. Némethy, and H. A. Scheraga, *J. Phys. Chem.*, **92**, 4739–4742 (1988).
62. IUPAC-IUB Commission on Biochemical Nomenclature, *Biochemistry*, **9**, 3471–3479 (1970).
63. D. M. Gay, *ACM Trans. Math. Software*, **9**, 503–524 (1983).
64. D. R. Ripoll, M. S. Pottle, K. D. Gibson, H. A. Scheraga, and A. Liwo, *J. Comput. Chem.*, **16**, 1153–1163 (1995).
65. G. Perrot and B. Maigret, *J. Mol. Graph.*, **8**, 141–148 (1990).
66. G. Perrot, B. Cheng, K. D. Gibson, J. Vila, K. A. Palmer, A. Nayeem, B. Maigret, and H. A. Scheraga, *J. Comput. Chem.*, **13**, 1–11 (1992).
67. M. L. Connolly, *J. Appl. Cryst.*, **16**, 548–558 (1983).
68. R. J. Wawak, K. D. Gibson, and H. A. Scheraga, *J. Math. Chem.*, **15**, 207–232 (1994).
69. W. Kabsch, *Acta Cryst. A*, **32**, 922–923 (1976).
70. W. Kabsch, *Acta Cryst. A*, **34**, 827–828 (1978).
71. R. Srinivasan and G. D. Rose, *Prot. Struct. Funct. Genet.*, **22**, 81–99 (1995).